Hochschule für Angewandte Wissenschaften Hamburg
Fakultät Wirtschaft und Soziales
Department Soziale Arbeit
Bachelor Soziale Arbeit

KI und Soziale Arbeit

Das Spannungsverhältnis zwischen Blackbox-Problematik und Reflexionsprozessen

Bachelor-Thesis

Tag der Abgabe: 25.09.2023

Vorgelegt von: Karl Classen
[redacted]

Betreuende Prüferin: Prof. Dr. Bettina Radeiski

Zweiter Prüfer: Prof. Dr. Peter Tiedeken

Gliederung

I. Abkürzungsverzeichnis

- AFST: Allegheny Family Screening Tool (In Pennsylvania angewandter Algorithmus zur Kindeswohlgefährdungseinschätzung)
- CPS: Child Protective Services (Für Kindeswohlgefährdung zuständige Behörde und Abkürzung für eine Kategorisierung von Kindeswohlgefährdungsmeldungen)
- DHS: Department of Human Services (US-Amerikanische Behörde verantwortlich für den sozialen Bereich)
- GPS: General Protective Services (Abteilung des DHS und Abkürzung für eine Kategorisierung von Kindeswohlgefährdungsfällen)
- KAIMo: Kann ein Algorithmus im Konflikt moralisch kalkulieren? (Leitfrage eines Forschungsteams in Deutschland und Name des dazugehörigen KI-Systems)
- KI: Künstliche Intelligenz (Maschinelles Datenverarbeitungssystem)
- KNN: Künstliches neuronales Netzwerk (Methode maschinellen Lernens)
- PP: Predictive Policing (Vorhersagetool, welches Anwendung in der Polizeiarbeit findet (oftmals KI-gestützt.))
- PRM: Predicitive Risk Modelling (Risikomodellierung, welche zur Einschätzung potenzieller zukünftiger Gefahren dient (oftmals KI-gestützt.))

1.Einleitung

In dieser Arbeit soll es um die Tiefe der Auswirkungen gehen, die digitale Technologien in Form von künstlicher Intelligenz auf das Denken von Sozialarbeiter:innen haben kann.
Der programmiertechnische Fortschritt der Datenverarbeitung hat innerhalb des letzten Jahres durch spektakuläre Neuerungen vor allem auch im Bereich der Sprachverarbeitung auf sich aufmerksam gemacht. Allen voran die Imitation menschlicher Sprachmuster durch Chatbots beeindruckte nach dem Erscheinen von “Chat GPT” die breitere Masse: So führte sie auch bei professionellen Journalist:innen eines Tech Portals zu der Erkenntnis, dass Technologie mittlerweile in der Lage ist, auch kompliziertere Textaufgaben wie das Verfassen eines Artikels kreativ lösen zu können (vgl. Bode 2023).
Wenn also die sprachlichen Kompetenzen einer Technologie mit der von Menschen vergleichbar werden, eröffnet sich ein neuer Horizont an Möglichkeiten in anderen potenziellen Arbeitsbereichen, in denen Sprache und Kommunikation eine zentrale Rolle spielen.
Die Rede ist hier vor allem von dem sozialen Bereich, in dem ethische Bedenken sich einen Kampf mit den verführerischen Potenzialen der Anwendung algorithmisierter Entscheidungshilfen liefern. Hierbei ist die Brisanz des Einsatzes der Technologie stark von dem genauen Anwendungsbereich abhängig: So wird der Einsatz von Chatbots zur Ausgabe von Informationen in der Migrationsberatung weniger stark diskutiert als ein digitales Assistenzsystem zur Einschätzung von Kindeswohlgefährdung.
Überall dort, wo die Technologie ihrer Funktion eines bloß ausführenden Werkzeugs entwächst und die Beeinflussung moralischer Entscheidungen für sich beansprucht, entsteht ein großes Konfliktfeld, in welchem die Ausmaße potenzieller Gefahren sowie möglicher Vorteile des Technologieeinsatzes diskutiert werden.

Beim Einsatz künstlicher Intelligenz im sozialen Bereich fällt vor allem das sogenannte “predictive risk modelling” durch eine sehr streitbare Inanspruchnahme zumindest eines Teils der Abschätzung sozialer Risiken auf. Das Entwicklungsziel solcher künstlichen Intelligenzen besteht nämlich darin, eine Vorhersage über die Entwicklung menschlichen Lebens zu treffen, wobei diese Vorhersage eine Prognose über das Gefahrenpotential für- oder ausgehend von den beurteilten Menschen beinhaltet.

Ein Blick auf die Funktionsweise der Technologie moderner künstlicher Intelligenzen verrät uns, dass das, was von informatischer Seite als intelligent bezeichnet wird, nur streckenweise mit menschlicher Intelligenz verglichen werden kann. Waren Algorithmen früher oftmals noch lange von Menschen geschriebene mathematische Formeln, die einem Computer die Verarbeitung von Daten Schritt für Schritt vorschrieben, so werden heutzutage selbstlernende Algorithmen durch den Einsatz riesiger Datenmengen (Big Data) darauf trainiert, eigenständig möglichst zutreffende Aussagen über die Realität zu generieren.
In Bezug auf dieses Thema taucht oftmals der Oberbegriff des maschinellen Lernens auf, welcher genau diesen Prozess bezeichnet. Beim sogenannten Lernen (Trainieren des Algorithmus') besteht eine populär gewordene Methode (Deep Learning, zu Deutsch Tiefes Lernen) darin, ein Netzwerk mehrerer Schichten von Verarbeitungseinheiten aufzusetzen, in welche die Programmierer:innen nur begrenzte Einsicht erhalten. Teile der Datenverarbeitung werden aufgrund fehlender Nachvollziehbarkeit der Rechenprozesse als Black Box bezeichnet.
Die Ausgabe des Algorithmus bewertet also menschliches Leben nach gewissen Kriterien, die zwar als Parameter in die Eingabeschicht eingeflochten werden, deren Gewichtung und Relevanz für die Ergebnisse der Berechnungen jedoch zumindest teilweise im Verborgenen bleiben.

An dieser Stelle ergeben sich zwei entscheidende Fragestellungen: Wie können moralische Entscheidungen durch eine Instanz gefällt werden, in deren Berechnungen nur eine begrenzte Einsicht gewährt ist? Wie lassen sich der Umstand, dass der Einsatz tief lernender Technologien ein Black-Box-Problem mit sich bringt, mit dem Grundsatz professioneller Reflexion in der Sozialen Arbeit vereinbaren?
Um diese Fragen zu klären, werden zunächst einige informatische Grundbegriffe erklärt, die für die Bearbeitung der Fragestellung relevant erscheinen.
Hiernach widmet sich die folgende Arbeit zwei realen Projekten aus der Fachwelt, die sich mit dem Einsatz künstlicher Intelligenzen im Gebiet der Erkennung von-/dem Umgang mit Kindeswohlgefährdungen auseinandersetzen:

- Zum einen wird das sogenannte "Allegheny Family Screening Tool", eine KI, die schon seit 2016 in den USA zum Einsatz kommt und Fachkräften im Jugendamt durch die Einblendung von Risikoscores bei der Gefahreneinschätzung potentieller Kindeswohlgefährdungen helfen soll, unter die Lupe genommen.

- Als zweites soll das Deutsche Projekt KAIMo analysiert werden, welches sich mit der Fragestellung nach moralischer Kalkulationsfähigkeit eines Algorithmus' am Beispiel der Kindeswohlgefährdung beschäftigt, wobei hier Expert:innen an den Schnittstellen der Fachgebiete Sozialer Arbeit, Philosophie und Informatik zusammenarbeiten.

Anschließend wird versucht, eine Übersicht zentraler Problemfelder, Debatten und Streitgegenstände rund um das Thema "Einsatz Künstlicher Intelligenz" im sozialen Bereich zu erstellen, um einen tieferen Einblick in unser Oberthema zu erhalten und die eigene Fragestellung besser im Gesamtdiskurs verorten zu können. Hierbei werden immer wieder Bezüge zu den beiden zuvor präsentierten KI-Projekten hergestellt, die immer wieder Rückschlüsse auf unsere Leitfragen beinhalten.

Zum Abschluss wird noch einmal auf die obenstehenden Leitfragen eingegangen und ein umfassendes Fazit gezogen.

2. Begriffsklärungen

Um im Folgenden die Funktionsweise der KI-Systeme besser verstehen zu können, ist es notwendig, zunächst mehrere zentrale Begriffe zu definieren.

Hierzu werden zunächst einige Fachbegriffe erklärt, die in der späteren Bearbeitung erneut aufgegriffen werden:

2.1 Künstliche Intelligenz

Der Begriff „Künstliche Intelligenz" entstand in den 50er Jahren im Rahmen der Entwicklung erster programmierbarer Computer (Paaß 2020, S.10). Zu der Zeit wurde in der Wissenschaft die bis heute immer wieder neu aufbereitete Frage diskutiert, zur Lösung welcher Probleme/Bearbeitung welcher (Teil-)aufgaben künstliche Intelligenz im Stande sei (vgl. Heinrichs et al. 2022, S.1).

Eine einheitliche Definition von Künstlicher Intelligenz existiert zurzeit nicht, jedoch gibt es verschiedene Annäherungen an den der KI innewohnenden Begriff der Intelligenz: Alpaydin stellt im Rahmen seiner Arbeit zum Maschinellem Lernen große Anforderungen an die Kompetenzen der Technik: „Um von Intelligenz sprechen zu können, muss ein System in einer sich verändernden Umwelt in der Lage sein [sic] zu lernen." (Alpaydin 2022, S.3).

Aber auch schon Mitte des letzten Jahrhunderts erzeugten neu entwickelte Computersysteme eine hohe Erwartungshaltung, welche unter anderem durch den

sogenannten Turing-Test - entwickelt vom britischen Mathematiker Alan Turing - auf den Prüfstand gestellt werden sollten (vgl. Paaß 2020, S.2). Der Test besteht im Kern darin, dass ein:e menschliche:r Schiedsrichter:in gleichzeitig mit einem sich in einem anderem Raum befindlichen Menschen und einer Maschine chattet, wobei der:die Schiedsrichter:in anhand von ihm:ihr gestellten Fragen herausfinden soll, wer von Beiden eine reale Person und wer eine Maschine ist (vgl. ebd.). Dieser Test, der in seiner Ursprungsform auf den Nachweis authentischer Nachahmung menschlichen Handelns zielt (vgl. Kirste 2019, S.21) kann analog auch für andere Bereiche wie zum Beispiel Erkennung von Bildinhalten erweitert werden (vgl. Paaß 2020, S.2). Kann die Leistung des Computers nicht von menschlicher Leistung unterschieden werden, wird dem Computer Intelligenz auf menschlichem Niveau attestiert (vgl. ebd.).
Das Konzept des Turing-Tests steht bei der Entwicklung intelligenter Computersysteme symptomatisch für die starke Orientierung an menschlicher Intelligenz im Fachbereich KI (vgl. u.A. Paaß 2020, S.1-2; Heinrichs et al. 2022, S.16; Mayer 2018, S.30).
Eine weitere Gemeinsamkeit der diversen KI-Forschung findet sich in Bezug auf die Funktionalität des zu entwickelnden Systems, welche die unterschiedlichen Ansätze zur Definition und Entwicklung künstlicher Intelligenzen eint: Nach Kirste und Schürholz „lässt sich [...] ein zentraler Aspekt benennen, den alle als KI bezeichnete [sic] Systeme aufweisen: Es ist der Versuch, ein System zu entwickeln, das eigenständig komplexe Probleme bearbeiten kann.“ (ebd. 2019, S.21)
Als ein Gradmesser der Eigenständigkeit von KI-Systemen dient die Unterteilung Dieser in zwei Kategorien:

2.2 Schwache KI

Als schwache KI gelten Systeme, die zur Erfüllung spezifischer Aufgaben dienen (vgl. Lehmann et al. 2021, S.19). Diese Systeme können auch als Fachidioten bezeichnet werden, da ihre Fähigkeiten, Probleme zu lösen, nicht über einen bestimmten Anwendungsbereich hinausgehen bzw. eine adäquate Reaktion auf unvorhergesehene Situationen nicht möglich ist (vgl. Kirste 2019, S.100).
Sie werden daher als Werkzeuge für spezifische Aufgaben benutzt. Innerhalb dieser Bereiche können Systeme als intelligent bezeichnet werden, welche zur Bewältigung von Aufgaben im Stande sind, für deren Bearbeitung ansonsten menschliche Intelligenz von Nöten wäre (vgl.

Lehmann et al. 2021, S.18). Innerhalb dieses Aufgabenbereichs kann die Lösungskompetenz einer schwachen KI die eines Menschen durch schnellere und/oder präzisere Bearbeitung übertreffen (vgl. ebd., S.19).
Was eine schwache KI limitiert, ist die Abhängigkeit der Systeme von ihren Programmierer:innen (vgl. Paaß et al. 2020, S.8-9). Sie ist abhängig davon, dass alle in sie einfließenden Informationen in eine für sie verarbeitbare Form gepresst werden (vgl. ebd. S.13). Es müssen Modelle entwickelt werden, die die reale Welt in für sie rechenbare Parameter darstellen, damit die KI einen Ausgabewert errechnen kann (vgl. ebd.). Wenn sich nun in der Wirklichkeit Umstände verändern, die dazu führen, dass das Modell keine oder nur noch bedingte Gültigkeit besitzt, müssen beispielsweise Parameter angepasst oder Klassifizierungen geändert werden (vgl. Vela et al. 2022, S.1). Das Phänomen, welches beschreibt, dass sich eine Umwelt so verändert, dass ein Modell zumindest Qualitätsverluste erleidet, nennt sich *„concept drift“* (vgl. ebd.).
Ein grundlegender Unterschied zum menschlichen Gehirn ist, dass eine schwache KI nur in beschränktem Maße auf sich verändernde Umwelten reagieren kann, weshalb eine flexible Anpassung an neue Situationen oft nicht möglich ist (vgl. Paaß et al. 2020, S.39).

2.3 Starke KI

Im Gegensatz zur schwachen KI „[versteht man unter] starker KI (engl. Strong AI oder Artificial General Intelligence) [hingegen Systeme], die über eine universelle Intelligenz verfügen“ (Lehmann et al. 2021, S.19).
Diese KI ist also in der Lage dazu, ein tieferes Verständnis von fachübergreifenden Themen zu entwickeln. Ebenso unterscheidet sie sich von schwacher KI dadurch, dass sie sich von selbst weiterentwickelt und dabei nicht auf die Bereitstellung von Trainingsdaten oder Konfigurierung der Funktionsweise durch Programmierer:innen angewiesen ist, sondern eigenständig lernen kann (vgl. IBM o.J. (b)).
Eine der Kompetenzanforderungen an starke KI ist demnach an den flexiblen Problemlösestrategien eines Menschen orientiert (vgl. Paaß et al. 2020, S.39). Eine solche KI ist jedoch derzeit noch nicht entwickelt worden.
Allerdings ist zu erwähnen, dass sich die Einsatzmöglichkeiten schwacher KI aufgrund des höheren Angebots von Daten sowie steigender Prozessorleistung in jüngerer Zeit schnell erweitern (Alpaydin 2022, Vorwort I).

2.4 Algorithmus

Doch wo liegen die Ursprünge hoher Anforderungen an eine KI? Der Begriff des Algorithmus kann uns hier weiterhelfen: Algorithmen basieren auf Modellen der Realität und sind mathematisch gesehen Formeln, die erstellt werden, um eingegebene Parameter zu einem Ergebnis zu verrechnen (vgl. Paaß 2020, S.50). Hierbei wird eine Schritt-für-Schritt-Anleitung befolgt, die zur Bearbeitung gewisser Aufgaben dient. „Die Absicht dahinter ist, einen Problemlöseprozess zu automatisieren […].“ (Lenzen 2020; zit. n. Steiner 2022, S.466)
Als Algorithmen werden zunächst also Anleitungen bezeichnet, von denen wir bei Angabe der geforderten Parameter ein gewünschtes Ergebnis erhalten. Eingegebene Informationen werden zu einem bestimmten Ergebnis weiterverarbeitet. Auch die Anleitung zum Zusammenbauen eines Ikea-Regals kann als Algorithmus bezeichnet werden, da auch hier die Befolgung eines jeden Schrittes zum gewünschten Ergebnis (Aufbau des Regals) führen wird, soweit alle Voraussetzungen wie zum Beispiel das Vorliegen des entsprechenden Werkzeugs erfüllt sind (vgl. Paaß 2020, S.50).
In der Welt der Informatik gilt also, dass für jeden Prozess, dessen Bearbeitung in einer Formel fassbar ist, ein Algorithmus programmiert werden kann.

Deduktive KI-Systeme setzen dies um, indem Regeln zur Verarbeitung eingegebener Daten implementiert werden (vgl. Lehmann et al. 2021, S.19). Hier besteht das Ziel darin, allgemeingültige Regeln in Programmiersprache zu definieren, die zur Ableitung logischer Schlüsse für eine spezifische Situation im Rahmen des Anwendungsbereichs geeignet sind (vgl. ebd.). So würde ein Algorithmus zur Berechnung des Nettogehalts einer:eines Angestellten das Bruttogehalt der Person in einer Formel mit allen obligatorischen Versicherungs- und Steuerabgaben verrechnen (vgl. Paaß 2020 S.50).
Wie aber kann ein Algorithmus Aufgaben lösen, deren genaue Bearbeitungsschritte uns nicht bekannt sind, bzw. nur abstrakt formuliert werden können?

2.5 Maschinelles Lernen

Auf der Suche nach der Antwort auf diese Frage entstand die Methode des maschinellen Lernens (vgl. 3Blue1Brown[1] 2017, 0:07-1:15). Hier existiert das populäre Beispiel der Klassifizierung digitalisierter handgeschriebener Ziffern in der Auflösung von 28x28 Pixeln (vgl. Paaß 2020, S.56; 3Blue1Brown[1] 2017, 3:10-3:20). Da wegen der unterschiedlichen Darstellungs- bzw. Schreibweise von Ziffern durch menschliche Hand nur sehr schwer allgemeine Regeln zur Zuordnung einer Zeichnung zu einer bestimmten Ziffer zu formulieren sind, werden solche Herausforderungen heutzutage mit Konzepten des maschinellen Lernens bewältigt (vgl. 3Blue1Brown[1]2017). „Die Idee des maschinellen Lernens besteht darin, das System lernen zu lassen, wie es solche Aufgaben löst." (Alpaydin 2022, S.1)
Für die sogenannten *induktiven KI-Systeme* wird nur eine grundlegende Struktur entworfen, welche sich dann durch weiteres Trainieren mit großen Datenmengen feinjustiert (vgl. Mayer 2018, S.31).

2.5.1 Künstliche neuronale Netzwerke (KNNs)

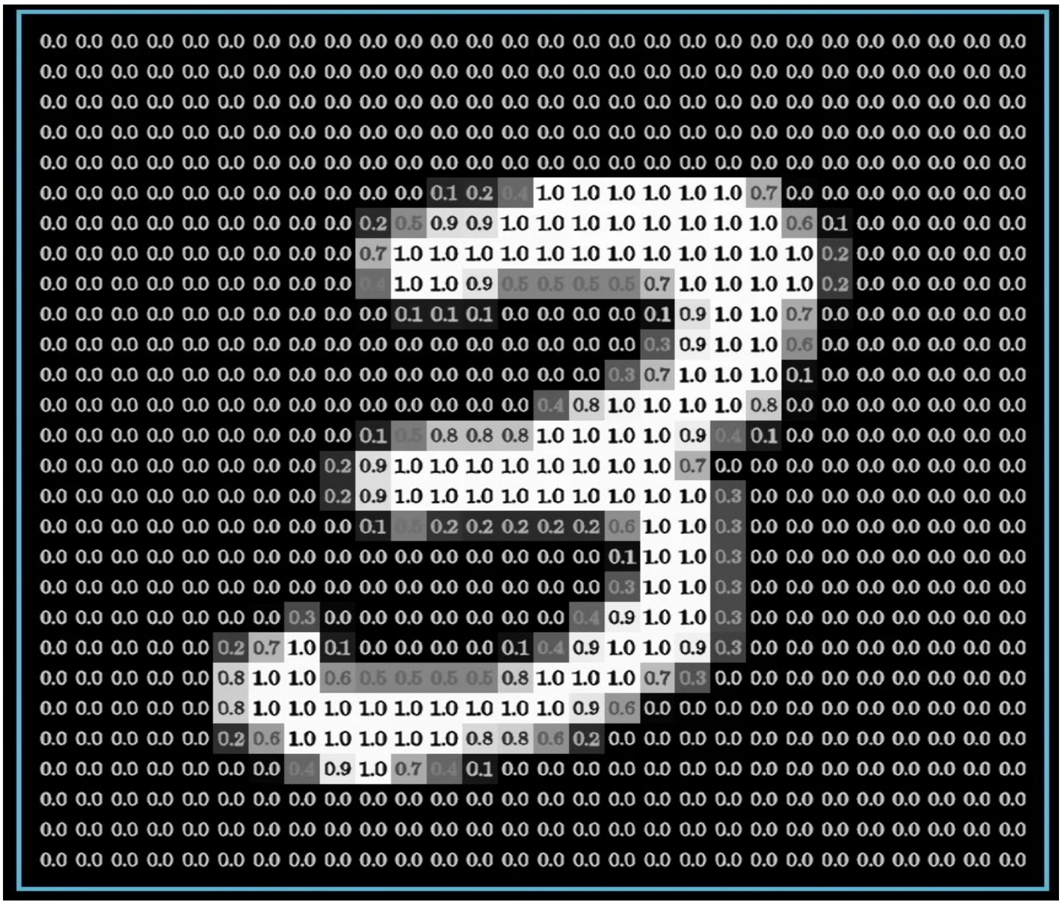
Ein berühmtes Beispiel für maschinelles Lernen ist die Methodik *künstlicher neuronaler Netzwerke (KNN)*, bei der sich das Training vereinfacht dargestellt folgendermaßen beschreiben lässt:
Ein KNN besteht aus mehreren Schichten künstlicher Neuronen beginnend mit der Eingabeschicht, in welche die uns bekannten Daten eingepflegt und in numerische Vektoren übersetzt werden und endend mit der Ausgabeschicht, welche uns das Ergebnis der Berechnung liefert (vgl. Lehmann et al. 2021, S.20). Um ein paar Einzelheiten besser nachvollziehbar zu machen, widmen wir uns an dieser Stelle der groben Umschreibung eines Beispiels für KNNs:
Genauer gesagt einem Netzwerk zur Klassifizierung handgeschriebener Ziffern - dargestellt in 28x28 Pixeln.

[1] Der Creator des Youtube Channels „3Blue1Brown", Grant Sanderson,(https://ocw.mit.edu/courses/18-s191-introduction-to-computational-thinking-fall-2020/) benutzt keine Quellenverweise, hat jedoch Mathematik und Informatik in Harvard studiert und leitete 2020 den Kurs "Introduction to computional thinking" am Massachusetts Institute for Technology als Co-Professor (https://ocw.mit.edu/courses/18-s191-introduction-to-computational-thinking-fall-2020/)
/(https://www.youtube.com/watch?v=vxjRWtWoD_w&list=PLP8iPy9hna6Q2Kr16aWPOKE0dz9OnsnIJ&t=82s)

Die Eingabeschicht könnte hier aus 784 Neuronen - dem Ergebnis von 28 mal 28 – bestehen. Nun würden in der Eingabeschicht die Verfärbungen eines jeden Pixels in Zahlenwerte gefasst und jeweils einem Neuron zugeordnet werden (vgl. 3Blue1Brown[1]2017, 3:10-3:31). Da die Bilder bei diesem Beispiel in Schwarz-Weiß dargestellt werden, kann die Verfärbung eines Pixels auf einer Skala von „0,0“ bis „1,0“ dargestellt werden (siehe Abb. 1). Hierbei steht der Wert „0,0“ für ein reines Schwarz und der Wert „1,0“ für einen vollkommen weiß gefärbten Pixel. Die dazwischenliegenden Werte fassen die unterschiedlich ausgeprägten Grautöne.

Die Ausgabeschicht besteht in diesem Fall aus 10 Neuronen, welche uns das Ergebnis der Klassifizierung liefern sollen – welche Ziffer von 0 bis 9 ist auf dem in die Eingabeschicht eingeflossenen Bild dargestellt?

Das Prinzip eines neuronalen Netzwerks fußt darauf, sogenannte versteckte Schichten (hidden layers) zwischen die eben erklärte Eingabe- und die uns das Ergebnis liefernde



Ausgabeschicht zu schalten (vgl. Lehmann et al. 2021, S.20).

Abbildung 1; Quelle: 3Blue1Brown[1] (2017): But what is a neural network? | Chapter 1, Deep learning

[YouTube]. Online unter: https://www.youtube.com/watch?v=aircAruvnKk (Zugriff: 11.06.2023)

Die genaue Größe und Anzahl der versteckten Neuronenschichten sind nicht an eine feste Bedingung geknüpft, wie es bei der Ein- und Ausgabeschicht, die abhängig von Pixelanzahl und Klassifizierungsmöglichkeiten sind, der Fall ist. Es besteht allerdings der Zusammenhang, je größer und zahlreicher die versteckten Neuronenschichten, desto zuverlässiger in der Regel das Ergebnis und desto höher die benötigte Rechenleistung (vgl. 3Blue1Brown[1] 2017).

In unserem Beispiel gibt es zwei versteckte Schichten a' 16 Neuronen. Jedes der ersten 16 Neuronen ist zur linken Seite mit jedem der 784 Neuronen aus der Eingabeschicht verknüpft. Auf rechter Seite ist jedes Neuron der ersten versteckten Schicht mit jedem Neuron der zweiten versteckten Schicht verknüpft. In der zweiten versteckten Schicht ist auf der rechten Seite jedes der 16 Neuronen mit jedem der zehn Neuronen der Ausgabeschicht verknüpft (siehe Abb. 2).

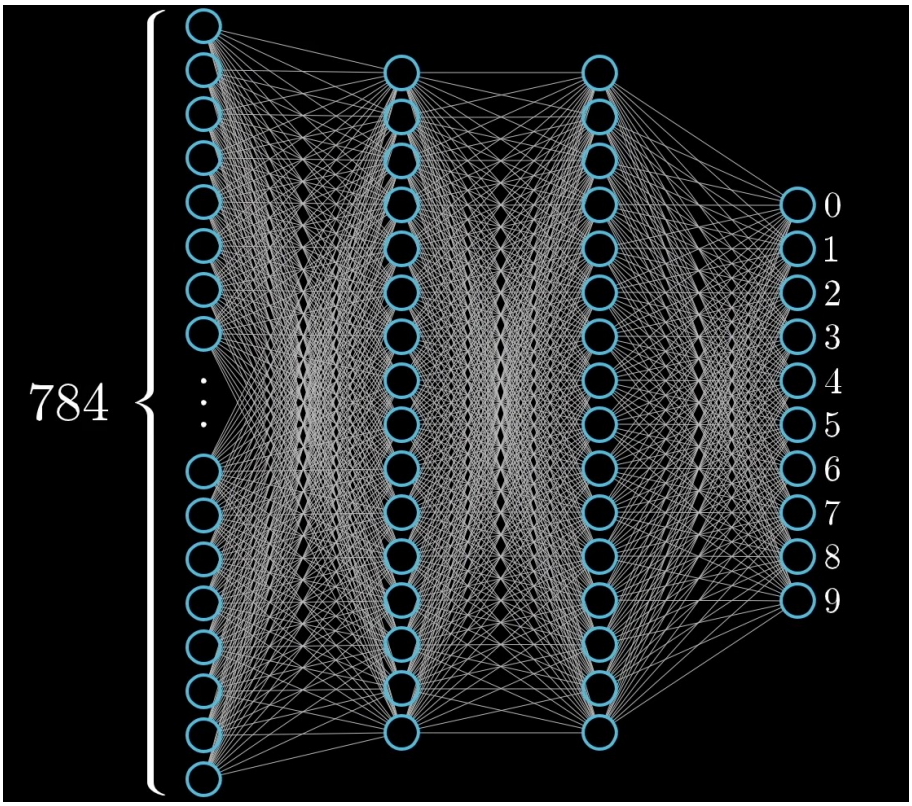


Abbildung 2; Quelle: 3Blue1Brown[1] (2017): But what is a neural network? | Chapter 1, Deep learning [YouTube]. Online unter: https://www.youtube.com/watch?v=aircAruvnKk (Zugriff: 11.06.2023)

Nun zur Verknüpfung: Unter Verknüpfung verstehen wir hier eine Form mathematischer

Abhängigkeit, die zwischen allen direkt verknüpften Neuronen besteht (vgl. ebd., 11:41-12:03). Ein neuronales Netz arbeitet hier mit sogenannten Gewichtungen und Aktivierungsgraden (vgl. ebd., 9:05-9:30).
In der Eingabeschicht würde die Skala des Aktivierungsgrads der Neuronen zwischen „0,0“ und „1,0“ liegen.
Nehmen wir das oberste Neuron der ersten versteckten Schicht als Beispiel:
Es ist mit allen 784 Neuronen der Eingabeschicht verknüpft. Für alle der 784 Verbindungen existiert ein Faktor (Gewichtung), der mit dem Aktivierungsgrad (Wert zwischen „0,0“ und „1,0“) aus der ersten Schicht multipliziert wird. Nun können diese Faktoren (Gewichtungen) positiv, negativ oder null sein. Die Summe all dieser 784 Multiplikationen bestimmt den Aktivierungsgrad unseres Neurons (vgl. ebd., 9:10-9:35).
Auf weitere mathematische Details kann an dieser Stelle verzichtet werden. Wichtig zu wissen ist aber, dass durch die Anpassung dieser Gewichtungen Informationen gewonnen werden können.
Würden zum Beispiel die Verbindungen zu Neuronen, welche die Pixel aus der Bildmitte repräsentieren, durch hohe Faktoren positiv gewichtet, wäre der Aktivierungsgrad unseres Neurons hoch, sobald sich in der Bildmitte viel weiße Farbe befände, da diese durch Werte nahe „1,0“ dargestellt wird (vgl. ebd., 9:40-9:55).
Eine weitere Möglichkeit zur Erkennung von Mustern, wäre eine Abgrenzung: Falls wir erkennen wollen, ob sich unser weißer Bildmittelpunkt durch schwarze Pixel von anderen Bildteilen abgrenzt, könnten wir für die Pixel um unseren positiv markierten Bereich negative Gewichtungen festlegen (vgl. ebd., 10:00-10:12).
Würden sich nun weitere weiße Pixel um den gefragten Bereich befinden, würde der Aktivierungsgrad unseres Neurons in der Summe sinken.
Dieses Neuron würde uns die Information liefern, ob sich im Bildmittelpunkt ein abgegrenzter weißer Fleck befindet (vgl. ebd.).
Zusammengefasst kann also jedes Neuron unserer beiden versteckten Schichten dazu genutzt werden, Muster zu erkennen. Zur Informationsverarbeitung werden also die unterschiedlich gewichteten Verbindungen zwischen den einzelnen Neuronen genutzt. In Abbildung 3 ist ein solches Verbindungsnetzwerk durch rote Striche für negativ gewichtete Verknüpfungen und grün für positive Gewichtungen veranschaulicht. Kräftigere Farben stehen dabei für stärkere Gewichtungen.

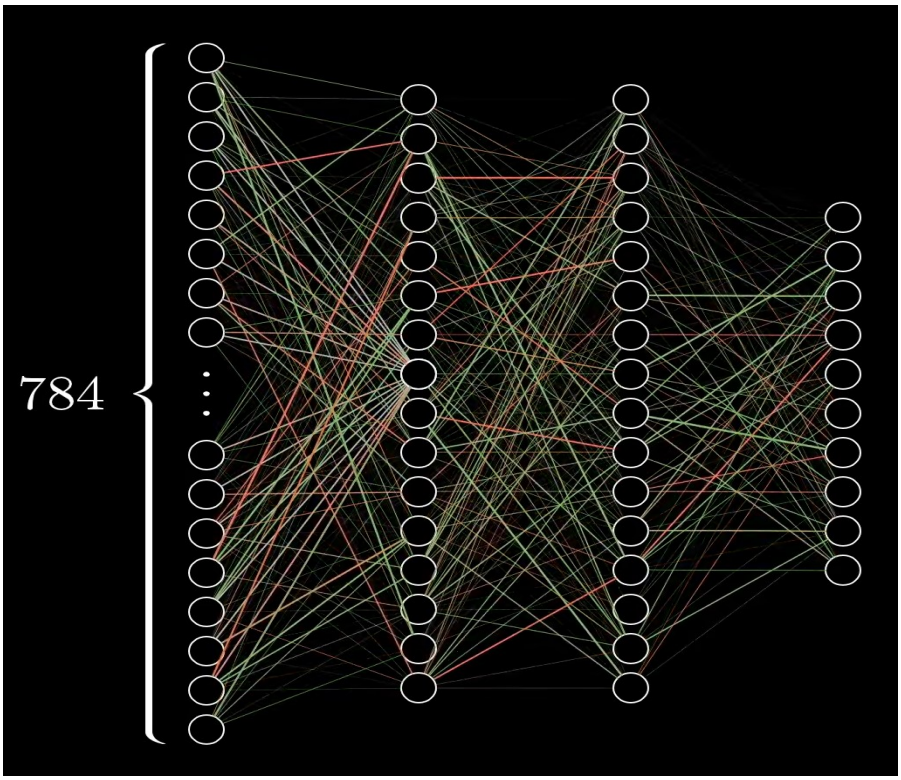


Abbildung 3; Quelle: 3Blue1Brown[1] (2017): But what is a neural network? | Chapter 1, Deep learning [YouTube]. Online unter: https://www.youtube.com/watch?v=aircAruvnKk (Zugriff: 11.06.2023)

Auch wenn die komplexen Mechanismen eines KNNs hier deutlich vereinfacht dargestellt wurden, ist dies das Grundprinzip nach dem ein künstliches neuronales Netzwerk – in diesem Fall zur Klassifizierung handgeschriebener Ziffern - arbeitet. Die Klassifizierung wird schlussendlich durch die unterschiedlichen Aktivierungsgrade der letzten Neuronenschicht (Ausgabeschicht) vorgenommen (Neuron mit höchstem Aktivierungsgrad = wahrscheinlichstes Ergebnis) (vgl. ebd., 15:28-15.43).
In der Realität werden die Gewichtungen zwischen allen Neuronen allerdings nicht von Hand angepasst, sondern eigenständig vom neuronalen Netzwerk im Training mit großen Datenmengen justiert:

> „Jeder Lernprozess verschiebt die Gewichtungen der Verbindungen in dem geschichteten Netzwerk. Die Aktivitätsbeziehungen zwischen den neuronalen Schichten ordnen sich, ohne dass ein Programmierer sie gezielt dekretiert. Die Entscheidungsstruktur verändert sich allein auf der Grundlage der verarbeiteten Daten.“ (Martini 2019, S.43)

Dies hat den Vorteil, dass sich kein Mensch hinsetzen muss, um - auf unser Beispiel bezogen - tausende Gewichtungen von Hand einzustellen.
„Beim Training lernen die inneren Schichten abstrakte Repräsentationen und grundlegende Eigenschaften." (Lehmann et al. 2021, S.20) So erarbeitet das KNN selbständig Wege, um die in die Eingabeschicht eingeflossenen Daten zu klassifizieren.
Ein Nachteil dieser Methode besteht allerdings darin, dass selbst für Programmierende solcher Algorithmen genaue Zusammenhänge der Ergebniskonstruktion nicht erklärbar sind (vgl. Martini 2019, S.43). Dies bezieht sich auf ein Fehlen von Nachvollziehbarkeit in Bezug auf die Gewichtungen der Verbindungen zwischen den einzelnen Neuronen. Die versteckten Schichten des Netzwerks bilden eine Blackbox, deren Verknüpfungen wir zwar einsehen, jedoch wenn überhaupt nur bedingt verstehen können (vgl. ebd.).

Innerhalb des maschinellen Lernens wird zwischen verschiedenen Lernverfahren unterschieden, von denen hier zwei Oberkategorien kurz erläutert werden sollen:

2.5.2 Überwachtes Lernen

Beim überwachten Lernen wird ein Algorithmus, beispielsweise ein künstliches neuronales Netzwerk, darauf trainiert, eigene Ausgabewerte den in den Trainingsdaten beinhalteten korrekten Ergebnissen möglichst genau anzunähern (vgl. Alpaydin 2022, S.10).
Ein Beispiel hierfür ist die Beschreibung realweltlicher Zusammenhänge mittels eines Modells - einer mathematischen Funktion. Hier können Parameter einer sogenannten Regressionsfunktion durch einen Algorithmus so optimiert werden, dass sich das Modell der Wirklichkeit bestmöglich annähert (vgl. ebd.).
Regression ist ein u.a. in der Sozialwissenschaft eingesetztes Verfahren, welches beispielsweise dazu dient, Einflussfaktoren (mathematisch: Parameter) auf bestimmte soziale Zustände oder Verhaltensweisen zu ermitteln und deren jeweilige Relevanz zu gewichten (vgl. Stoetzer 2017, S.7-9). Diese Einflussfaktoren werden in der Sozialwissenschaft Prädiktoren oder unabhängige Variablen genannt, deren Auswirkung auf die sogenannten abhängigen Variablen/Zielvariablen untersucht wird (vgl. Ng et al. 2018, S.73).
Das Entdecken dieser Zusammenhänge mittels Regressionsanalysen wird in der Praxis oft zur

Erstellung von Prognosen genutzt (vgl. ebd. S.6).
Auch *Klassifizierung* von Text- oder Bilddateien kann mittels überwachten Lernens gelingen (vgl. Alpaydin 2022, S. 10). Auch hier gibt der Algorithmus stets die laut eigenen Berechnungen wahrscheinlichste Klassifizierungsoption an, welche mit dem vorgegebenen Ergebnis der Trainingsdaten abgeglichen wird. Dies führt im Laufe des Trainings zur bestmöglichen Anpassung der für die Klassifizierung zuständigen sogenannten Diskriminanzfunktion (vgl. ebd.).
Diese Funktion passt sich in einer Art und Weise an, die es dem System ermöglicht, das Klassifizieren zu lernen.

2.5.3 Unüberwachtes Lernen

„Unüberwachte Lernverfahren (Unsupervised Learning) versuchen ohne menschlichen Input, latent in den Daten vorhandene Strukturen aufzudecken." (Lehmann et al. 2021, S.19)
Da von menschlicher Seite keine Vorgaben für die Produktion von Informationen gemacht werden, wie zum Beispiel Angabe möglicher Klassifizierungsoptionen, besteht der Anwendungsbereich des unüberwachten Lernens darin, den Algortithmus selbständig nach Mustern in Trainingsdaten suchen zu lassen und diese möglichst trennscharf zu klassifizieren (vgl. ebd.).
Beim unüberwachten Lernen werden dem Algorithmus keine korrekten Ausgabewerte zum Abgleich eigener Berechnungen zur Verfügung gestellt (vgl. Alpaydin 2021, S.11-12).
Die sogenannte *Clusteranalyse* ist eine Methode des unüberwachten Lernens, deren Ziel darin besteht, „Cluster (Häufungen) oder Gruppierungen von Eingabewerten zu finden." (ebd. S.12)
Die Ordnungsstruktur wird also vom Algorithmus selbst erarbeitet, sodass Muster in großen Datenmengen schnell erkannt werden können.

3.1 Fallbeispiel 1: Allegheny family screening tool

Kommen wir nun zum konkreten Anwendungsbereich KI-gestützer Technologien:
Zunächst beschäftigen wir uns mit einem im Allegheny County, zu Deutsch Allegheny Landkreis, welcher die engere Metropolregion um die Stadt Pittsburgh fasst, angewandten algorithmischen Assistenzsystem.
Hierbei handelt es sich um das „Allegheny Family Screening Tool" (AFST), einem Algorithmus

zur Risikomodellierung von Kindeswohlverdachtsfällen, der seit 2016 als Assistenzsystem in dem „department of human services“ (DHS) im Bereich „child welfare“ in Allegheny eingesetzt wird (vgl. Allegheny County (a) 2023)
Auf das deutsche System bezogen ist der für unsere Untersuchung relevante Teil der Behörde vergleichbar mit einem Jugendamt.

3.1.1 Anwendungsbereich des Algorithmus

Das AFST wurde für den Zweck entwickelt, die „Call Service“-Kräfte in Alleghenys Behörde bei eingehenden Meldungen über Kindeswohlgefährdungen zu unterstützen.
Um sich den Einsatzbereich der Servicekräfte in der Rufbereitschaft genauer vorstellen zu können, folgt nun ein kurzer Überblick über die Zuständigkeit der Mitarbeitenden des DHS, denen das AFST als Assistenzsystem bereitgestellt wird:
In Pennsylvania werden Meldungen zu potenziell kindeswohlgefährdenden Ereignissen oder Zuständen in zwei Kategorien aufgesplittet.
Um das Einsatzgebiets des Algorithmus abzustecken, folgt nun eine kurze Übersicht der Fälle erster Kategorie, die nicht in unseren Anwendungsbereich fallen:
Fälle erster Kategorie werden als „CPS“ klassifiziert, hierzu gehören Meldungen, deren Inhalt der Definition von „child abuse“, zu Deutsch Kindesmissbrauch, entspechen (vgl. Vaithianathan et al. 2017, S.6). Für diese muss innerhalb von 24 Stunden eine Untersuchung vom „Child Protection Service“ (CPS) eingeleitet werden (vgl ebd.).
Wenn aus der Meldung also hervorgeht, dass ein Kindesmissbrauch nach Pennsylvanischer Gesetzgebung vorliegt, wird keine weitere Risikoabschätzung der Fachkräfte vorgenommen, sondern innerhalb eines Tages gehandelt, in dringenden Fällen auch sofort (vgl. Rubin 2020, S.14).
Die Definition von Kindesmissbrauch findet sich im Child Protective Services Law (CPSL) und umfasst die folgenden Voraussetzungen:
Einem Individuum unter 18 Jahren wurde durch ein Handeln oder nicht Handeln, welches innerhalb der letzten zwei Jahre stattgefunden hat, Schaden zugefügt oder es wurde eben dadurch gefährdet(vgl. Rizvi et al. 2023, S.3-4). Der:die Täter:in muss dabei zumindest leichtfertig gehandelt haben, in dem Sinne, dass „[ein nicht zu rechtfertigendes, substantielles Risiko bewusster Weise missachtet wurde.]“ (Rizvi et al. 2023, S.4 – Übersetzung von mir, KC).

Praktische Beispiele für Straftaten, die unter Kindesmissbrauch fallen, wären Anwendung sexueller Gewalt oder das Hervorrufen schwerwiegender seelischer Verletzungen bei dem:der betroffenen Minderjährigen (vgl. Rizvi et al. 2023, S.2).
Die Meldungen der zweiten Kategorie sind die sogenannten GPS-Fälle (vgl ebd.). GPS steht für „General Protective Service“ und Klassifizierungen dieser Art ziehen keine umgehende Untersuchung nach sich, erfordern jedoch die Beurteilung einer Fachkraft, über möglicherweise zu ergreifende Maßnahmen (vgl. ebd.).
Die Schwelle für GPS-Fälle ist niedriger als die für CPS-Fälle: „[…] eine Verweisung wird als GPS-Meldung kategorisiert, wenn Dienstleistungen notwendig sein könnten, um einen potenziellen Schaden, welcher bestimmte Voraussetzungen erfüllt, für ein Kind zu verhindern.” (Rubin 2020, S.14 – Übersetzung von mir, KC)
Im Gegensatz zur CPS-Fall-Voraussetzung „Kindesmissbrauch“ ist hier die Rede von potenziellem Schaden für das Kind – die Gefährdungslage ist demnach weniger drastisch und weniger eindeutig. Aber auch sie wird bei der Erfüllung bestimmter Voraussetzung genauer ermittelt.
Diese „bestimmten Voraussetzungen“ sind durch die Angabe von insgesamt 32 GPS-Fall-Subkategorien durch das Department of Human Services (DHS) von Pennsylvania konkreter beschrieben (vgl. ebd. S.9-13).
Zur Einordnung des Schadensbegriffs sind hier einmal drei GPS-Subkategorien gelistet:

- Unentschuldigter Schulabsentismus für mindestens drei Tage pro Schuljahr
- Freiwilliger Drogen- und/oder Alkoholkonsum des Kindes
- Ein allein gelassenes Kind ohne Sorgeberechtigten

Für die weitere Gefahreneinschätzung einer als GPS kategorisierten Meldungen kommt nun das Allegheny Family Screening Tool (AFST) zum Einsatz (vgl. Vaithianathan et al. 2017, S.6). Aus diesem Grund wird sich im Folgenden nicht auf CPS-, sondern nur auf GPS-Fälle bezogen.

3.1.2 Funktionsweise des AFST

Das AFST ist ein Assistenzsystem, welches dafür entwickelt wurde, das Personal der Hotline für Kindeswohlgefährdungsmeldungen bei der Entscheidung zu unterstützen, ob beim vorliegenden Fall eine weitere Gefahreneinschätzung vor Ort notwendig ist (vgl. ebd. S.4).
Bei dem System handelt es sich um ein „predictive risk model“ (PRM), es wird eingesetzt, um zukünftige Risiken zu modellieren und dadurch mögliche Gefahren für Minderjährige besser

abschätzen zu können (vgl. ebd., S.8).
Das Tool soll es den Mitarbeitenden ermöglichen, den Hauptfokus ihrer Arbeit auf die Beschreibungen der eingetroffenen Meldung zu legen, während mehr als 100 Prädiktoren (Parameter) in ein Regressionsmodell zur Berechnung eines Risikoscores einfließen (vgl. Dalton 2022, S.1; Allegheny County DHS 2019, S.5; Vaithianathan et al. 2019, S.3).
Der Algorithmus berechnet pro Minderjährigen einen Risikoscore zwischen 1 und 20, der für jedes Kind vorhersagen soll, wie hoch die Wahrscheinlichkeit ist, innerhalb der nächsten zwei Jahre in einer Jugendwohnung, Pflegeeinrichtung o.Ä. untergebracht zu werden (vgl. Vaithianathan et al. 2020, S.3). Hierbei differenziert der Algorithmus bei Meldungen, die mehrere Kinder beinhalten - beispielsweise wegen gleicher Haushaltszugehörigkeit - nicht zwischen den Risikoscores der Beteiligten, sondern gibt der Call-Screener-Kraft den höchsten der verfügbaren Scores an (vgl. Allegheny County DHS 2019, S.5).
Der:die Call-Screener:in erhält also neben der Beschreibung der spezifischen Situation, die in der Meldung enthalten ist, eine Evaluation des allgemeinen Gefährdungsgrades der betroffenen Minderjährigen.
Wichtig zu erwähnen ist noch, dass der Risikoscore ausschließlich zur ersten Gefahreinschätzung durch Hotline-Servicekräfte verwendet wird und den Fachkräften, die vor Ort mit den Kindern und Familien arbeiten, nicht mitgeteilt wird (vgl. Vaithianathan et al. 2017, S.32).
Das „Department of Human Services“ von Allegheny County betont ebenso, dass die Entscheidung, weitere Ermittlungen zu veranlassen, nicht von den „call screeners“ getroffen werden, sondern hierfür die sogenannten „screening supervisors“ verantwortlich sind (vgl. Allegheny County DHS 2019, S.10). Sie werden von den „call screeners“, die alle Infos zum Fall zusammentragen, gebrieft und müssen im Anschluss eine Entscheidung treffen, wobei der Risikoscore nur als einer von mehreren Indikatoren dienen soll (vgl. ebd.).
Doch wie berechnet sich dieser Score? Was sind Prädiktoren, die die Zielvariable (Risikoscore) beeinflussen?

3.1.3 Berechnung des Risikoscores

Zuallererst ist zu erwähnen, dass es zwei Versionen des AFST gibt, da das System im November 2018 ein Update erhielt (Allegheny County DHS 2019, S.4). Im Folgenden wird das Augenmerk vor allem auf die neuere Version (Version 2) gerichtet, wobei allerdings auch

Bezug auf das ursprüngliche Modell (Version 1), dessen Weiterentwicklung und Anpassung genommen wird. Wird Version 1 nachfolgend nicht explizit erwähnt, ist in den anschließenden Erläuterungen Version 2 gemeint.
Zur Berechnung des Risikoscores werden mehr als 100 Prädiktoren verwendet (vgl. ebd. S.5). Diese Prädiktoren werden mit personenbezogenen Daten gefüttert, die in unserem Fall aus 21 verschiedenen Quellen des integrierten Datensystems des DHS stammen können (vgl. ebd. S.6; Vaithianathan et al. 2019, S.3). Das DHS verfügt über Zugriff auf das „data warehouse“ von Allegheny County, in dem eine große Bandbreite von Daten gesichert werden (vgl. Vaithianathan 2019, S.3)
Inkludiert in dem Datensystem sind zu großen Teilen auch Informationen über Leistungen des DHS selbst - zum Beispiel „child protective services“, öffentlich geförderte Programme für psychische Gesundheit oder Arbeit mit suchterkrankten Menschen (vgl. Allegheny County DHS 2023; Allegheny County DHS 2019, S.6). Daten aus verschiedensten Sektoren des Sozialsystems werden demnach zentral durch das „Department of Human Services“ verwaltet und stehen dem AFST zur Berechnung des Risikoscores zur Verfügung.
Auf informatischer Ebene passiert vereinfacht gesagt Folgendes:
Für die Entwicklung der ersten Version des AFST wurden 76.964 GPS-/CPS-Meldungen zwischen April 2010 und April 2014 als Trainingsdaten benutzt (vgl. Vaithianathan 2017, S.11). Nun wurde für jede der Meldungen festgestellt, ob in dem der Meldung nachfolgenden Zeitraum die in der Meldung erwähnten Kinder/Jugendlichen in einer Pflegeeinrichtung („foster care placement“) o.Ä. untergebracht wurden (vgl. ebd.). Das Prinzip von PRM ist die Ermittlung von Einflussfaktoren, die das Eintreten des Risikos („foster care placement“) begünstigen:
Mithilfe eines Regressionsmodells sollten hier also Zusammenhänge zwischen „foster care placement“ und der sozialen Situation des Kindes -modelliert durch Informationen aus dem Datensystem des DHS, eingespeist als unabhängige Variablen (Prädiktoren) - hergestellt werden (vgl ebd., S.11-13).

3.1.4 Prädiktoren

Werfen wir nun einen genaueren Blick auf die einzelnen Prädiktoren, die die Ausgabe des Algorithmus – also den Risikoscore – beeinflussen:
Wie oben bereits erwähnt, fließen über 100 Variablen in die Berechnungen des Algorithmus

ein, weshalb es den Rahmen sprengen würde, diese hier detailliert aufzulisten, allerdings lohnt es sich einen beispielhaften Blick auf einige der Prädiktoren zu werfen.
Der Algorithmus strukturiert fast alle Eingaben nach Personen, die in der GPS-Meldung Erwähnung finden oder Sorgeberechtigte:r/Elternteil des Kindes sind, für welches der Risikoscore erstellt wird (vgl. Vaithianathan et al. 2019, S.16).
Dabei existieren neben den Sorgeberechtigten, dem:der angeblichen Täter:in (welche:r Deckungsgleich mit anderen Kategorien sein kann) und dem mutmaßlichen Opfer noch die Kategorie weiterer mutmaßlicher Opfer, für die in dieser Berechnung kein Risikoscore erstellt wird, und die Kategorie „andere Kinder, die in dem GPS-Report vorkommen (vgl. ebd.). Menschen aus den eben erwähnten Kategorien werden im Folgenden teils als „beteiligte Personen“ zusammengefasst, womit nicht zwangsweise alle, aber immer mehr als eine Kategorie gemeint ist.
Ganz oben auf der Listung der Prädiktoren durch die Entwickler:innen des AFST findet sich die Variable „Alter“ des Kindes, welches laut Meldung mutmaßliches Opfer ist (vgl. ebd.). Hier unterteilt der Algorithmus die Kinder in grobe Altersgruppen, anstatt ein genaues Alter zu erfassen (vgl. ebd.). Die Altersspanne vergrößert sich dabei innerhalb der Altersgruppen mit steigendem Alter der Betroffenen, sodass 13- bis 18-Jährige in eine Kategorie zählen, während am anderen Ende des Spektrums noch zwischen null- und einjährigen Kindern unterschieden wird (vgl. ebd.).
Dies könnte aus entwicklungspsychologischer Perspektive Sinn ergeben, da Ereignisse innerhalb des ersten Lebensjahres große Auswirkungen auf die weitere Entwicklung eines Kindes haben und Kinder in frühen Entwicklungsphasen besonders abhängig von ihren Versorger:innen sind (vgl. Kasten 2014, S.3-4).
Eine gröbere Unterteilung im Jugendalter aufgrund steigender Unabhängigkeit der Minderjährigen wäre hier denkbar (vgl. Lohaus 2018, S.26).
Die Variablen beziehen allerdings nicht nur Informationen über die Kinder, sondern auch über deren näheres Umfeld mit ein (vgl. Vaithianathan et al. 2019, S.16).
So werden zum Beispiel sozioökonomische Faktoren wie der schulische/akademische Abschluss der Mutter oder die sogenannte „poverty rate“ - per Definition ein Instrument zur Messung von Armut einer Gruppe im Vergleich zum Medianhaushalt einer Obergruppe (vgl. OECD 2023) – als Prädiktoren gelistet (vgl. Vaithianathan et al. 2019, S.19-20).
Des Weiteren spielen psychosoziale Faktoren eine Rolle: So fließt die Information, ob eine

der beteiligten Personen eine Historie im „behavioral health“-System vorweist, mit in die Informationsverarbeitung des AFST ein (ebd. S.20). „Behavioral health“, zu Deutsch Verhaltensgesundheit, bezieht sich auf Angebote für psychisch erkrankte- oder alkohol-/drogenabhängige Personen (Allegheny County (b) 2023)
Andere Prädiktoren erfassen die Gefängnishistorie der beteiligten Personen:
So wird die Dauer des Aufenthalts innerhalb der letzten ein bis drei Jahre im Allegheny County Jail gemessen, indem für jeden Zeitraum, die im Strafvollzug verbrachten Monate gezählt werden (vgl. Vaithianathan et al. 2019, S.20).
Für vor kürzerer Zeit (letzte drei Jahre) kriminell oder auffällig gewordene Jugendliche aus der GPS-Meldung werden Daten aus den „Juvenile Probation-Records“ abgerufen (vgl. ebd.). „Juvenile Probation“ ist die in US-Amerikanischen Jugendgerichten am häufigsten verhängte Sanktion (vgl. Annie E. Casey Foundation 2021). Es handelt sich dabei um ein Spektrum an möglichen Interventionen, welche die Jugendlichen oftmals dazu zwingen, sich an gewisse Regeln wie zum Beispiel eine Ausgangssperre oder Umgangsverbot mit gewissen Personenkreisen zu halten (vgl. ebd.). Die Jugendlichen stehen dabei unter der Aufsicht eines Bewährungshelfers, mit dem sie regelmäßig Kontakt halten müssen (vgl. ebd.).
„Juvenile Probation“ wird allerdings nicht nur kriminellen Jugendlichen verordnet, sondern ist auch ein Kontrollinstrument für beispielsweise durch Schulabsentismus auffällig gewordene Jugendliche (vgl. ebd.).

3.1.5 Veränderungen des Algorithmus und „concept drift“

Im November 2018 erschien ein Update der im August 2016 gestarteten ersten Version des AFST, mit dem Veränderungen bezüglich verwendeter Prädiktoren und deren Modellierung sowie Anpassungen des Einsatzgebietes des Algorithmus und neue Visualisierung der Ausgabewerte des AFST einhergingen (vgl. Vaithianathan 2019, S.2-5). Nachfolgend wird der Fokus auf den Bereich der einfließenden Prädiktoren begrenzt.
Der Einbezug armutsbezogener Daten der AFST V2 sank im Vergleich zur ersten Version des AFST. So bezieht die zweite Version keine Daten aus „public benefit records“, zu Deutsch etwa staatliche Wohlfahrtsprogramme, mit in die Berechnungen ein, da sich Daten über Leistungsbezüge hier mit der Zeit so gewandelt hatten, dass Klassifizierungen nicht mehr mit den Trainingsdaten in Einklang standen (vgl. ebd., S.3-4). Ursprünglich wurden beispielsweise Daten des „Supplemental Nutrition Assistance Programm“ (SNAP) oder der

„Temporary Aid to Needy Families“ (TANF) – beides Indikatoren für niedrige finanzielle Ressourcen – verwendet (vgl. ebd. S.3).
Ebenso wurden aufgrund von Gesetzesänderungen Klassifizierungen im Gesundheitssystem so angepasst, dass Details zu „behavioral health records“ in der zweiten Version nicht mehr angemessen verarbeitet werden konnten (vgl. ebd. S.4).
Was hat es mit der Schwierigkeit, den Algorithmus an veränderte Klassifizierungssysteme anzupassen, auf sich?
Wie oben unter dem Punkt „schwache KI“ bereits erwähnt, kann ein Algorithmus schwer auf spontane Veränderungen seines Bezugssystems (Umwelt) reagieren. In diesem Fall wurde das AFST (V1) mit Daten, die zwischen April 2010 und April 2014 erhoben wurden, trainiert (vgl. Vaithianathan 2017, S.11).
Was den Bereich „behavioral health records“ angeht, bezog sich ein Prädiktor der ersten Version zum Beispiel auf die Klassifizierung „neurotic disorder“ (vgl. ebd. S. 36).
Da laut Entwickler:innen des AFST „behavioral health“-Diagnosen neu kategorisiert und definiert wurden (vgl. Vaithianathan 2019, S.4), könnte also unter Menschen mit „neurotic disorder“ seit den Neuerungen eine andere Personengruppe gefasst sein oder aber Diagnosen wurden unter neuen Namen anders geordnet. Für unseren Algorithmus heißt das, dass der Prädiktor, welcher sich auf die Diagnose „neurotic disorder“ bezieht, nicht mehr verwendet werden kann, da die Diagnose nicht mehr existiert oder da nun eine andere Personengruppe als die, mit denen in den Trainingsdaten gelernt wurde, unter diesen Begriff fällt.
Das veränderte Klassifizierungssystem der „behavioral health“-Diagnosen kann als „concept drift“ bezeichnet werden, auf welchen eine schwache KI nicht eigenständig reagieren kann. Das Modell muss also von den Programmier:innen umstrukturiert werden und in einem überwachten Lernprozess mit den neuen Klassifizierungen der „behavioral health-records“ trainiert werden, wenn genauere Diagnosen wieder in die Berechnungen des AFST mit einfließen sollen (vgl. ebd. 2019, S.4).

3.2 KAIMo

Kommen wir nun zum deutschen Ansatz zur algorithmischen Kalkulation von Risiken für das Kindeswohl. Das Forschungsprojekt KAIMo - finanziert von dem Bayerischen Institut für Digitale Transformation (bidt) – wird geleitet von den drei Professoren Michael Reder

(Professor für Praktische Philosophie), Nicholas Müller (Professor für Soziоinformatik) und Robert Lehmann (Professor für Soziale Arbeit) (vgl. Jaskolla o.J., 0:10-0:57).
In Bezug auf die Entwicklung des Systems messen die Forschenden der Anpassung des Algorithmus an die bisherige Arbeitsstruktur der Fachkräfte besondere Bedeutung bei, um letztendlich Arbeitsaufwand zu verringern und somit Entlastung zu schaffen (vgl. Burghardt o.J., 19:39-19:57).
Daher ist ein kurzer Blick auf den späteren Anwendungsbereich des Algorithmus obligatorisch.

3.2.1 Anwendungsgebiet von KAIMo

Wie reagiert das Jugendamt in Deutschland auf eine Kindeswohlgefährdungsmeldung? Zunächst ist festzustellen, dass das Jugendamt in Deutschland Kindeswohlgefährdungsmeldungen nicht in zwei Kategorien einteilt, sondern nach einer Handlungsvorschrift auf Basis von §8a SGB VIII agiert (vgl. Müller 2018, S.294):

> „Werden dem Jugendamt gewichtige Anhaltspunkte für die Gefährdung des Wohls eines Kindes oder Jugendlichen bekannt, so hat es das Gefährdungsrisiko im Zusammenwirken mehrerer Fachkräfte einzuschätzen."

Das Bekanntwerden gewichtiger Anhaltspunkte kann zum Beispiel durch Hinweise (Anrufe, Mails etc.) von Privatpersonen oder Institutionen wie der Schule oder des Kindergartens erfolgen, allerdings können auch Fachkräfte des Jugendamtes im Zuge ihrer Arbeit Anzeichen einer möglichen Kindeswohlgefährdung entdecken (vgl. Burghardt o.J., 4:15-4:53).
Werden solche gewichtigen Anhaltspunkte Mitarbeitenden des Jugendamtes bekannt, ist dieses laut §8a Abs. 1 Nr.1 SGB VIII dazu verpflichtet, sich „einen unmittelbaren Eindruck von dem Kind und von seiner persönlichen Umgebung […] zu verschaffen", falls dies erforderlich erscheint. Mögliche Interventionen finden sich im Gesetzestext – die drastischste aller Maßnahmen ist unter Absatz 2 gelistet:

> „Besteht eine dringende Gefahr und kann die Entscheidung des Gerichts nicht abgewartet werden, so ist das Jugendamt verpflichtet, das Kind oder den Jugendlichen in Obhut zu nehmen."

Zusammengefasst sind Mitarbeitende des Jugendamtes demnach verpflichtet, Hinweise auf mögliche Kindeswohlgefährdungen zu prüfen, diese im fachlichen Diskurs genauer einzuschätzen und falls notwendig Schutzmaßnahmen wie die Inobhutnahme des Kindes zu ergreifen (vgl. Burghardt o.J., 3:49-4:05)
Die Fachkräfte haben Entscheidungshoheit über den genauen Ablauf des Prozesses der Untersuchung, sind jedoch einerseits an fachliche Standards, wie die Kontrolle der Bedürfnisbefriedigung betroffener Kinder oder die Einschätzung elterlicher Erziehungskompetenz (vgl. Alle 2012, S.57), sowie andererseits an die Berücksichtigung rechtlicher Güter wie dem Recht auf Erziehung seitens der Eltern und der Schutzverpflichtung des Jugendamtes gegenüber dem Kind als Teil des Wächteramtes gebunden (vgl. Müller 2018, S.302-303).

3.2.2 KAIMo, ein Forschungsprojekt

Zur Unterstützung der Fachkräfte, die innerhalb dieses komplexen Systems arbeiten, versuchen Mitarbeitende der Hochschule für Philosophie München, der Technischen Hochschule Nürnberg sowie der Hochschule für Angewandte Wissenschaften Würzburg-Schweinfurt ein algorithmisches Assistenzsystem zu entwerfen (vgl.Hochschule für Philosophie München o.J.). Hierbei steht die namensgebende Leitfrage „Kann ein Algorithmus im Konflikt moralisch kalkulieren?“, (ebd.) im Zentrum des Erkenntnisinteresses. Geprüft werden soll, „ob institutionelles Handeln in moralischen Konfliktfällen durch Softwareprogramme digital unterstützt oder gar ersetzt werden kann“ (ebd.).
Im Gegensatz zum AFST ist KAIMo ein Forschungsprojekt (vgl. ebd.) und findet daher (noch) keine Anwendung im praktischen Bereich. Ebenso ist die verfügbare Literatur zu dem Projekt im Vergleich zum AFST in Stückzahl und Breite deutlich limitierter. Die Analyse des Projektes kann sich also im Folgenden nicht auf eine konkrete Arbeitsweise des Algorithmus beziehen, sondern orientiert sich vor allem an den von den Entwickler:innen herausgegebenen Fachtexten, Statements und den dort enthaltenen Beschreibungen zu der KI.

3.2.3 Drei Wissenschaften

Zuerst ist zu sagen, dass das Projekt aus einem Zusammenschluss von Akademiker:innen aus den drei Fachrichtungen Philosophie, Informatik und Soziale Arbeit besteht (vgl. Reder o.J.).

Neben der Synthese aus Sozialer Arbeit und Informatik wird hier also Wert auf eine philosophische Perspektive der Dinge gelegt. Rebecca Gutwaldt – philosophische Mitarbeiterin des Projekts – sieht eine Kernaufgabe ihrer Profession darin, ethische Konfliktherde im Bereich des Kinderschutzes zu identifizieren, um eine Reflexion des Systems unter moralischen Gesichtspunkten zu fördern (vgl. Gutwaldt o.J., 1:30-1:45).
Maximilian Kraus ist wissenschaftlicher Mitarbeiter für Sozioinformatik – damit Bestandteil des Fachteams für Informatik in KAIMo - und beschäftigt sich mit einer informatischen Umsetzung des Projektes, die sinnvoll für die sozialarbeiterische Praxis sein soll (Wie können sozialarbeiterische Zusammenhänge technisch adäquat operationalisiert werden?) (vgl. Kraus o.J., 1:58-2:18).
Jennifer Burghardt – ebenfalls wissenschaftliche Mitarbeiterin des Projekts - beschreibt als zentrale Aufgabe aus sozialarbeiterischer Perspektive, einen empirischen Wirkungsnachweis der eingesetzten Technik zu erbringen (vgl. Burghardt o.J., 2:17-2:45). Ihr geht es also darum, nachweisen zu können, dass durch den Einsatz von KI die Vorhersagequalität bezüglich Kindeswohlgefährdungen steigt (vgl. ebd.).

3.2.4 Philosophische Herangehensweise

Um die Funktionsweise von KAIMo besser erläutern zu können, lohnt es sich, zunächst einen kurzen Blick auf die dem Projekt inhärenten philosophischen Grundanschauungen in Bezug auf ethische KI zu werfen: Michael Reder – verantwortlich für die philosophische Leitung des Projekts (vgl.Reder o.J.) – und Christopher Koska – wissenschaftlicher Mitarbeiter am Lehrstuhl für praktische Philosophie und ebenfalls Projektkoordinator von KAIMo (vgl. ebd.) – stellen ihre KI als „artificial moral agent“ vor und grenzen die Fähigkeiten künstlicher Intelligenz bezüglich moralischer Entscheidungsfindung anhand der Dimensionen des *Kalkulierens*, *Abwägens*, *Entscheidens* und *Handelns* ein (vgl. Koska et al. 2023, S.2).
Hier wird ferner zwischen drei moralischen Konflikten unterschieden, von denen laut Koska und Reder zwei nicht in den Zuständigkeitsbereich von „artificial moral agents“ fallen (vgl. ebd. S.3): Einerseits das moralische Dilemma, welches dann vorliegt, „wenn zwei moralische Forderungen sich gegenseitig ausschließen und nicht beide gleichzeitig verfolgt werden können.“ (ebd.) Und andererseits auflösbare moralische Konflikte, die durch Zuführung neuer Informationen oder Klärung von Missverständnissen und Kommunikationsfehlern aufgelöst werden können (vgl. ebd.).

Im Zentrum stehen moralisch gewichtige Konflikte, welche zwar entscheidbar sind, jedoch nicht durch die Behebung eines Problems, sondern vielmehr durch die genaue Abwägung moralischer Forderungen und Güter (vgl. ebd. S.4). Die Beurteilung von Gefährdungslagen Minderjähriger, die zu einem Eingriff in das elterliche Fürsorgerecht zum Wohle des Kindes führen kann, kann unter diese Definition subsumiert werden.
Nun zurück zu den vier Dimensionen, die im Folgenden auf die Arbeit des Jugendamtes im Rahmen von Kindeswohlgefährdungseinschätzungen bezogen werden: Das konkrete *Entscheiden* in Konfliktfällen über den endgültigen Verbleib des Kindes bleibt Gerichten überlassen, die in ihr Urteil konkrete Lebensumstände der Beteiligten sowie die Gewichtung verschiedener Rechtsgüter einbeziehen müssen (vgl. ebd. S.4).
Letztlich wird auch die Dimension des *Handelns* in moralischen Kontexten als Fähigkeit kategorisiert, die vorerst Menschen vorbehalten bleibt, da Maschinen nicht über personale Selbstbestimmung verfügen und somit keine eigenen moralischen Standpunkte vertreten können (vgl. ebd.).
Was sie hingegen können, ist die *Abwägung* moralischer Güter, die vor den Prozessen des Handelns und Entscheidens erfolgt, zu begleiten (vgl. ebd.). „In einer eher technischen Semantik geht es um die moralische *Kalkulation*." (ebd., S.5)
Das Ziel ist, den Algorithmus so zu programmieren, dass beispielsweise Fallzusammenhänge durch das System geordnet dargestellt oder sonstig gesammelte Daten durch KAIMo konzeptualisiert werden können (vgl. ebd.). Hierdurch sollen Fachkräfte bei der Einschätzung des Falles, eine gehobene Übersicht gewinnen, was zu einer „Verbesserung der Einschätzung und Abwägung – eben der moralischen Kalkulation" (ebd.) – führen soll.
Wie sich die konkrete Unterstützung der Fachkräfte durch das Assistenzsystem genau gestaltet, wird im kommenden Unterpunkt genauer beleuchtet.

3.2.5 Funktionsweise von KAIMo

In dem aktuellsten wissenschaftlichen Artikel zum Projekt stellen Michael Reder und Christopher Koska zunächst nur grobe Ideen des KAIMo-Prototypen vor, weshalb in vielen Bereichen noch auf keine endgültige oder detailscharfe Vorgehensweise geschlossen werden kann, sich jedoch ein Umriss des Projekts abzeichnen lässt (vgl. ebd. S.11).
In dem Artikel offenbaren die Beiden die Verfolgung eines Ansatzes, der weniger auf die Erstellung einer konkreten Vorhersage in Form von Wahrscheinlichkeiten oder Scores zielt,

sondern legen den Schwerpunkt auf die Unterstützung der Reflexion der Mitarbeitenden in den unter Abbildung 4 zu sehenden Arbeitsschritten (vgl. ebd.).
Das Vorgehen deutscher Jugendämter kann laut dem Forschungsteam um KAIMo in drei Arbeitsschritte unterteilt werden (vgl. ebd. S.9). Um eine hilfreiche Unterstützung zu sein, soll KAIMo hieran adaptiert werden:
Wie in Abbildung 4 zu sehen ist, besteht das Assistenzsystem aus drei Komponenten, welche auf die drei Arbeitsschritte „Assessment", „Planning" und „Controlling" zugeschnitten sind.
Assessment: Unter dem Punkt „3.2.1 Anwendungsbereich von KAIMo" wurde bereits beschrieben, dass das eine Kindeswohlgefährdungseinschätzung initiierende Ereignis die Meldung einer Kindeswohlgefährdung bzw. ein aufkommender Verdacht einer Fachkraft ist. Bevor eine mögliche Intervention eingeleitet wird, wird nach Eingang einer solchen Meldung im Rahmen des Assessments die Sachlage des vorliegenden Falls in Zusammenarbeit mehrerer Fachkräfte evaluiert.
Planning: In dieser Phase wird das weitere Vorgehen auf die vorherige Assessment-Phase zugeschnitten und geplant. Fachkräfte diskutieren anhand dokumentierter Erkenntnisse aus der Arbeit mit den Familien, wie sie die Gefährdungslage einschätzen, ob Maßnahmen zu ergreifen sind und falls ja, welche.
Controlling: In dieser Phase wird mit Blick auf den weiteren Hilfeverlauf die Wirksamkeit der ergriffenen Maßnahmen überprüft.
In Abbildung 4 wird die Integration der drei Komponenten (Bots) von KAIMo in das Arbeitssystem des Jugendamtes visualisiert. Im Folgenden sollen Aufgaben, Funktionsweise und Systematik der Bots näher erläutert werden.

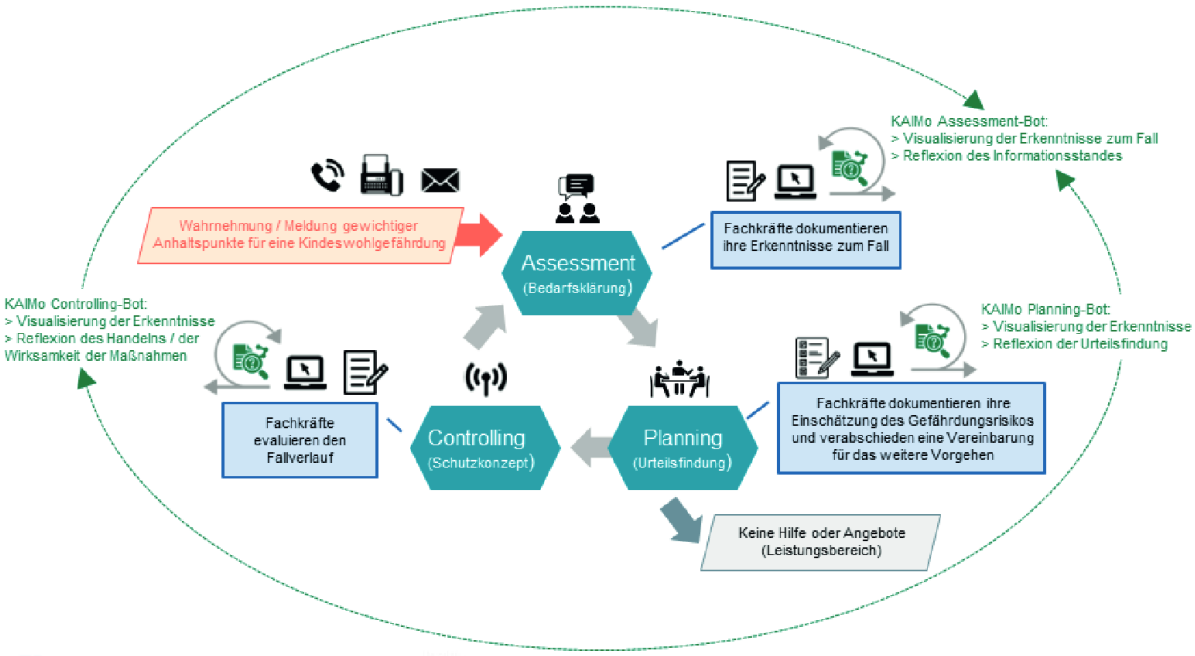


Abbildung 4; Quelle: Koska et al. 2023, S.10

Welche Datengrundlage befähigt die Bots dazu, Professionellen der Sozialen Arbeit Reflexionsunterstützung bieten zu können?
Die schnellste Antwort auf diese Frage ist einfach: Der Algorithmus soll durch die Bereitstellung von Falldaten trainiert werden (vgl. ebd. S.9). Bei genauerem Hinsehen können diese Daten allerdings in zwei Kategorien unterschieden werden: Zum einen gelang es dem KAIMo-Team, einige anonymisierte Fallakten aus Jugendämtern zu akquirieren, diese stellten jedoch keine ausreichende Datengrundlage dar (vgl. ebd.).
Deshalb griff das Team ergänzend auf einen alternativen Weg der Datenakquise zurück: Hierfür wurden fiktive Fallszenarien online gestellt und von Fachkräften der Sozialen Arbeit bewertet (vgl. ebd.). Es wurden also künstlich Daten aus ausgedachten, aber an realen Umständen angelehnten Fällen, erzeugt – sogenannte synthetische Daten (vgl. ebd.).
Eine Kombination beider Datentypen wurde zum Trainieren der KI verwendet.

3.2.5.1 Technische Werkzeuge

Durch welche technischen Werkzeuge werden diese Datensätze ausgewertet?
Die Datensätze werden durch Verfahren des Maschinellen Lernens aufbereitet, in diesem Fall durch „natural language processing“ (NLP) (vgl. ebd., S.10). NLP beschreibt die computergestützte Analyse von Textdaten, welche Textmaterial zum Beispiel nach

Satzbaustrukturen, semantischen Zusammenhängen oder sonstigen abstrakten Konzepten untersuchen soll (vgl. Chowdharry 2020, S.604). Eine Methode, die hier Anwendung findet heißt „multi-label-text-classification“ (vgl. Koska et al. 2023, S.10). Bei „Multi-label-text-classification“ handelt es sich um das Konzept, die KI Texte mit sogenannten „Labels“ markieren zu lassen, die helfen sollen, einen Text nach seinen Inhalten zu kategorisieren (vgl. Nam et al. 2014, S.437; Zhao 2023, S.1-2). Mit dieser Methode wäre es in unserem Fall zum Beispiel möglich, Fallakten nach bestimmten Überbegriffen zu sortieren, wodurch eine schnellere Übersicht erstellt und Gemeinsamkeiten, Unterschiede und/oder Zusammenhänge zwischen aktuell vorliegenden und archivierten Fällen schneller sichtbar gemacht werden könnten. Eine feinstufigere Verwendung von NLP wird als Basis aller drei Bots beschrieben (vgl. Koska et al. 2023, S.10):

In allen drei Bereichen des Systems sollen zusammengetragene Informationen zum Fall mithilfe eines Knowledge Graphen visualisiert werden (vgl. ebd.). Ein Knowledge Graph ähnelt von der Struktur her einer Mindmap und dient bezogen auf KAIMo der visuellen Strukturierung von Fallinhalten (vgl. IBM o.J. (b)).

Abbildung 4 deutet daraufhin, dass es sich bei dem im Knowledge-Graph festgehaltenen Wissen um vorgefilterte Erkenntnisse der Fachkräfte handelt, es ist aber anhand der Graphik und des dazugehörigen Artikels nicht zweifelsfrei zu bestimmen, ob auch maschinell verarbeitete Rohdaten (die ungefilterte Meldung) in das System einfließen. Ein Hinweis auf die ursprüngliche Intention, dies zu tun, findet sich in einem Interview, welches zu Zwecken der Weiterbildung Sozialarbeitender im Fachbereich KI aufgenommen wurde:

So benennen die Entwickler:innen dort als ein Ziel ihres Projekts, mit dem Assistenzsystem die Mitarbeitenden des Jugendamtes durch Verringerung des Arbeitsvolumens entlasten zu wollen (vgl. Burghardt o.J., 19:41-19:55). Eine konkrete Idee dazu war ursprünglich, eine beim Jugendamt als Text oder Audiodatei eingegangene Kindeswohlgefährdungsmeldung durch den Algorithmus dokumentieren und voranalysieren zu lassen, bevor sich menschliche Fachkräfte überhaupt mit der Meldung beschäftigen (vgl. Kraus o.J., 15:24-16:03). So sollte zwar auch Bias, der durch die Filterung der Meldung durch eine Fachkraft fast zwangsweise entstehe, abgebaut werden, allerdings sah das KAIMo-Team hierin ebenfalls Potenzial für Arbeitsersparnis (vgl. ebd., 16:03-16:23). Zusammengefasst bleibt ein wenig unklar, welche Informationen genau in die Aufbereitung durch die Bots einfließen.

Als nächstes soll geklärt werden, wie genau sich die Aufgaben der drei Bots differenzieren?

Welche Unterschiede gibt es in ihren Anwendungsbereichen?

3.2.5.2 Drei Bots

Assessment-Bot: Eine Aufgabe des Assessment-Bots wäre es, die im Knowledge Graphen zusammengetragenen Informationen aus dem aktuell zu bearbeitenden Fall auf Vollständigkeit zu prüfen, indem die Merkmalslage des gegenwärtigen Falls mit derer vergangener Szenarien verglichen wird (vgl. Koska et al. 2023, S.10). So könnten Arbeitskräfte auf Bereiche stoßen, die im Kontext dieses Falles noch unerforscht geblieben sind, eventuell aber erhöhte Relevanz besitzen. Zum Beispiel könnte dadurch auffallen, dass wenig Informationen zur Vater-Kind-Beziehung vorhanden sind, diese aber als durchaus fallrelevant gelten darf.

Planning-Bot: Während der kollegialen Beratung zum weiteren Vorgehen soll das System potenzielle Voreingenommenheiten der Mitarbeitenden einblenden, um eine möglichst objektive Planungsphase zu ermöglichen. Ebenfalls soll der Planning-Bot die Ressourcen der Mitarbeitenden auf die bedeutsamsten Aspekte des Falls bündeln:

> „In der Planungsphase wird vor allem durch Regressionsbildung der extrahierten Fallmerkmale zu Fällen des Trainingsdatensatzes eine Vorhersage getroffen: Welche Merkmale besitzen für den vorliegenden Fall die (vermeintlich) größte (fachliche) Aussagekraft und erfordern deshalb die größte Aufmerksamkeit der Fachkräfte?“ (ebd.)

In diesem Arbeitsschritt greift KAIMo also durchaus auf algorithmische Vorhersagetools zurück, allerdings lässt sich aus diesem Zitat nicht herauslesen, dass den Anwender:innen durch die Prognosen des Planning-Bots verschiedene Gefährdungsgrade in Bezug auf unterschiedliche Fälle suggeriert werden.

Vielmehr wird der Algorithmus durch Trainingsprozesse darauf programmiert, die Fachkräfte kontextsensibel auf Merkmale, die sich in vergleichbaren Fallkonstellationen als besonders ausschlaggebend herausgestellt haben, hinzuweisen.

Als potenzielles Werkzeug hierfür wird „bayesian linear regression“ gelistet (vgl. ebd.) – ein lineares Regressionsmodell mit einer Besonderheit:

So zielen die Bayes’schen Regressionsmethoden nicht allein darauf, Anwender:innen des Modells eine eindimensionale abhängige Variable, wie zum Beispiel den Risikoscore in Form einer konkreten Zahl, zu liefern, sondern beruhen darauf, eine Wahrscheinlichkeitsverteilung möglicher Ergebnisse auszugeben (vgl. Koehrsen 2018).

Dies lässt sich gut an einem Beispiel erläutern:
Eine Methode der Bayes'schen Regression – das „Bayesian Model Averaging" - verwendet bezüglich einer einzigen Berechnung, also zur Berechnung einer Ausgabe, mehrere Modelle gleichzeitig (vgl. Hinne et al. 2020, S.201-202). Diese Modelle könnten zum Beispiel alle auf (leicht) unterschiedlichen Annahmen aufgebaut sein, sodass sie zu unterschiedlichen Ergebnissen kommen. Durch diese Methode können also die Prognosen mehrerer Modelle verarbeitet werden, wobei die Prognosen eines jeden Modells entsprechend ihrer Probabilität unterschiedlich gewichtet werden (vgl. ebd., S.202). So fließen Ausgabewerte von Modellen, deren Vorhersagen sich in der Vergangenheit öfters als wahr oder nah an der Realität herausgestellt haben, stärker in die Vorhersage ein als solche deren Vorhersagen weniger Wahrheitsgehalt hatten. Trotz der Modelldiversität fußt diese Methode auf dem Integrieren mehrerer Modelle in ein großes Modell.
Wir erhalten demnach ein diversifizierteres Bild der Zukunft und können aus den Differenzen der verschiedenen Prognosen Aussagen bezüglich der Unsicherheit unserer Modelle ableiten. Wenn diese also alle sehr unterschiedliche Ergebnisse liefern würden, wäre eine Schlussfolgerung, dass Prognosen über die Zukunft als sehr unsicher gelten können. Bayes'sche Regression „erlaubt uns also, unsere Unsicherheit über das Modell zu quantifizieren" (Koehrsen 2018 – Übersetzung von mir – KC).
Anstatt einer auf eine Variable konzentrierte Ausgabe erhalten wir durch Bayes'sche Regression eher eine Art Wahrscheinlichkeitskorridor, der uns ebenso über das Unsicherheitsmaß der Schätzung informiert (vgl ebd.).
Doch nun zurück zur Vorgehensweise des Planning Bots: Nachdem eine Vorhersage über die (vermeintlich) aussagekräftigsten Merkmale des Falls getroffen wurde, sollen die Prognosen der KI durch die Fachkraft mithilfe eines Chatbots kritisch reflektiert werden (vgl. Koska et al. 2023, S.10). Eine Fachkraft des Jugendamtes würde sich also hinsetzen und die durch KAIMo produzierten Vorhersagen mit einer virtuellen Chatpartnerin, der KI, diskutieren. Eine mögliche Gesprächsführungsmethode ist hierbei der sokratische Dialog, ein Prinzip des Fragenstellens und Hinterfragens, welches das Ziel verfolgt, den:die Gesprächspartner:in durch offene Gesprächsführung in einem Prozess der Urteilsbildung und Erkenntnisgewinnung zu unterstützen (vgl. Heurer o.J.)
Controlling-Bot: In dem Text von Koska und Reder lässt sich nicht klar zwischen der anfänglichen Aufgabe des Controlling-Bots und der des Planning-Bots differenzieren. So

steht dort nur, dass „[im] dritten und letzten Schritt [kontextsensitive Reflexionsfragen gestellt werden], die aus dem aktuellen Sachstand und dem jeweiligen Arbeitsschritt der Fachkräfte erschlossen werden.“ (Koska et al. 2023, S.10)
Dies bezieht sich wohl auf den letzten Arbeitsschritt des Controllings und impliziert weitere Unterstützung der Reflexion durch Chatbots.

3.2.6 Zukunftspläne

Abseits der jetzigen Möglichkeiten finden ebenso Ideen für Zukunftspläne ihren Platz in dem Artikel. Der Hauptgrund dafür, dass diese im Prototypen noch keinen Platz finden, ist wohl der Mangel an umfassenderen Daten aus dem Arbeitsfeld. Diese sollen aber mit der Zeit durch Fachkräfte gesammelt werden und neue Funktionen ermöglichen:
Zunächst ziehen Koska und Reder einen Vergleich zwischen ihrer Systemanwendung und der von selbstfahrenden Autos (vgl. edd. S.8-9). Hier orientiert sich die Funktionsweise des Algorithmus an als effektiv bewerteten menschlichen Vorgehensweisen.
Im sogenannten Schattenmodus lernt ein Computersystem Autofahren, die KI steuert zunächst jedoch nicht das Fahrzeug, sondern immer noch ein Mensch. Das System sammelt jedoch währenddessen Daten, die KI-Systeme dann nach Mustern des menschlichen Fahrverhaltens untersuchen (vgl. ebd. S.9). Dieses Konzept soll dazu beitragen, „Szenarien-Kataloge für vergleichbare Unfall und Gefahrensituationen [zu] erstellen und die jeweils erfolgreichsten Lösungsstrategien für unterschiedliche Szenarien [zu] ermitteln.“ (ebd.)
Bevor der Algorithmus also selbst das Fahrzeug steuert, operiert er als stiller Lerner im Hintergrund. Die Datenressourcen des Algorithmus beschränken sich dabei nicht auf die Daten eines einzigen Autos, sondern werten Datenmengen riesiger Autoflotten aus (vgl. ebd., S.9-11). Die Fachterminologie hierfür ist „fleet learning“, zu Deutsch etwa Flottenlernen, und bezieht sich auf den kollektiven Wissensgewinn. Tesla ist eine Firma, die auf diese Art der Datenverarbeitung setzt: Ihre Wagen funktionieren als Netzwerk, innerhalb dessen aus Einzelfällen erlernte Lösungsstrategien auf die ganze Flotte übertragen werden (vgl. Strobl 2017). Die durch „fleet learning“ gesammelten Informationen bilden zusammengenommen einen riesigen Erfahrungsschatz, auf den jeder Teil des Netzwerks zurückgreifen kann.
Für KAIMo würde das bedeuten, dass die KI sich am Anfang noch im Hintergrund hält und im Rahmen des Controllings Daten aus verschiedensten Teilen der Flotte – in diesem Fall einem

Netzwerk aus Jugendämtern - über den tatsächlichen Weiterverlauf einer Hilfe sammelt, um so Zusammenhänge zwischen Merkmalslage eines Falls und Effektivitätsgrad der jeweiligen Hilfe zu erhalten (vgl Koska et al. 2023, S.10-11).
Laut Koska und Reder würde KAIMo also mit der Zeit lernen, welche Maßnahmen in welchen Fallkonstellationen effektiv wären (vgl. ebd., S.11). Das würde es dem System ermöglichen, nach der Anamnese in der Assessmentphase Mitarbeitenden konkrete Handlungsempfehlungen zu geben (vgl. ebd.). So würde sich dann der Kreis zwischen Controlling und Assesment schließen, da die aus beiden Phasen gewonnenen Daten zu einer Handlungsempfehlung weiterverarbeitet werden können.
Für die Auswertung der neu zu liefernden Daten aus den Jugendämtern soll „clustering“ ein Werkzeug sein (vgl ebd.), welches - wie bereits unter „2.5.3 Unüberwachtes Lernen“ erwähnt - zu den unüberwachten Lernmethoden gehört. Dies würde nahelegen, dass auch nicht standardisierte Daten aus der Anamnese durch den Algorithmus geordnet werden könnten, um diese thematisch zu kategorisieren, sodass eine Sortierung nach Merkmalslage der Fälle möglich wäre.
Zur Abschätzung der Folgen von Maßnahmen sind neuronale Netzwerke die Methode der Wahl (vgl. ebd.). Diese sind - wie unter dem Punkt „2.5.1 Künstliche neuronale Netzwerke (KNNs)“ bereits veranschaulicht - gut dafür geeignet, Muster in Daten zu erfassen und könnten hier Zusammenhänge zwischen Merkmalslage und ergriffenen Maßnahmen erkennen.
Da der Nutzen dieser Methoden sich mit dem Zweck der Anwendung Bayes’scher Regression überlagert, ist schwierig zu sagen, ob beide Verfahren zusammen funktionieren könnten oder ob „clustering“ und KNNs das Bayes’sche Regressionsverfahren mit der Zeit ablösen werden.

4. KI im sozialen Bereich

Nachdem nun zwei spezielle Systeme genauer beleuchtet wurden, widmen wir uns als Nächstes mit einem deutlich gröberen Blick dem Einsatz verschiedener KI-Technologien im sozialen Bereich. Damit ist in diesem Fall das Anwendungsgebiet gemeint, in dem Menschen in direktem Zusammenhang mit einer KI-Technologie stehen. Darunter fällt auch der Aspekt der Forschung im sozialen Bereich. Hier ist es zwar möglich, dass Personen, deren Daten durch Algorithmen gesammelt und aufbereitet werden in keiner direkten Verbindung zum

Forschungsprojekt stehen, jedoch werden auch hier Daten über menschliche Verhaltensweisen analysiert.
Nicht in den von uns untersuchten Bereich würde zum Beispiel in der Industrie eingesetzte Robotik mit integrierten KI-Systemen fallen, da hier zwar Menschen von der Technologie profitieren und diese auch entwerfen, der Zweck der KI sich allerdings rein auf das Fertigen eines Produktes beschränkt.

4.1 Big Data

Um zu verstehen, weshalb das Phänomen algorithmischer Anwendungen in den letzten Jahren so stark an Popularität gewonnen hat, kommt man ohne die Erklärung des Begriffs „Big Data“ nicht aus. Big Data meint eine große Menge an Daten, deren Produktion durch gesellschaftliche Digitalisierungsprozesse ermöglicht wurde (vgl. Gutwald et al. 2021, S.4). Im Zeitalter des Internets ist es so leicht wie noch nie, große Mengen an Daten zu erfassen und immer größere Datenmengen in riesigen Datenbanken zentral zusammenzutragen (vgl. Hasan 2013, S.29).
Dabei geschieht längst nicht jeder Datenproduktionsprozess im Internet freiwillig:
„Jede Aktivität hinterlässt digitale Datenspuren, teils sind es bewusst eingegebene Inhalte, teils unbewusst und technisch automatisch erzeugte Meta- und Beobachtungsdaten.“ (Gapski 2020, S.157)
Diese hohe Verfügbarkeit an Daten einerseits trifft andererseits auf neue technische Methoden zur Verarbeitung der Daten gepaart mit stetig steigender Prozessorleistung (vgl. Alpaydin 2022, Vorwort XV).
All die Daten, die beispielsweise durch Navigationssysteme, Social Media, Internetchroniken o.Ä. gewonnen werden, können vielfältig weiterverarbeitet werden und dienen in der Datenwirtschaft hauptsächlich zur Extraktion von Informationen über die nutzenden Personen (vgl. Ng et al. 2018, Vorwort VII). So könnten gewonnene Daten nach Mustern untersucht werden, die Aufschluss über das Kaufverhalten einer Person geben, um anschließend ein Kundenprofil zu erstellen und gezielt Werbung zu schalten (vgl. ebd.).
Das gängigste Mittel zur Analyse der Daten ist die Nutzung maschinellen Lernens (vgl. ebd., Vorwort VIII).
Kritiker:innen sehen in dem Wertgewinn privater Daten eine zunehmende Bedrohung der Privatsphäre von Individuen. So spricht Shoshana Zuboff in diesem Kontext vom

„Überwachungskapitalismus“, eine durch finanzielle Anreize angetriebene Form der Überwachung, welche mit privaten Erfahrungen als primärem Rohstoff dealt (vgl. 2019, S.5).

4.2 KI und Datenschutz

Wie oben bereits erwähnt, ist der Grad der Freiwilligkeit, der zur Preisgabe personenbezogener Daten führt, nicht immer gleich oder dem Subjekt ist die Erhebung dieser Daten zumindest nicht in vollem Umfang bewusst. Bei einer ordnungsgemäßen Einwilligung zur Verwendung von Cookies haben Privatunternehmen wie Google nämlich die Möglichkeit, diese Daten zur Erstellung von Persönlichkeitsprofilen zu nutzen (vgl. Google 2023; iRights.Lab 2017). Diese werden durch einen einzigen Klick akzeptiert und führen zur Erhebung von Metadaten durch Google, also Daten, die sich nicht auf intentional geteilte Inhalte beziehen, sondern im Rahmen der Nutzung erhoben, ergo im Hintergrund digitaler Aktivitäten gesammelt werden (vgl. Google 2023).
Hierunter befinden sich auch sehr sensible Daten wie der Standortverlauf des:der Nutzers:Nutzerin (vgl. ebd.). Diese werden dann unentwegt gespeichert und verarbeitet. Solche Entwicklungen lassen die Frage aufkommen, ob der durch künstlich intelligente Systeme gestiegene Anreiz, Daten zu verarbeiten, zur Verletzung ethischer und rechtlicher Normen wie der Achtung von Privatsphäre oder zur Einschränkung des Rechts auf informationale Selbstbestimmung führt.
Vorteile von automatisierter Datenauswertung beschränken sich allerdings nicht nur auf die Privatwirtschaft, sondern auch Forschende der Sozialwissenschaften sehen durchaus Nutzen in den Methoden von Big Data-Science:
Ein Beispiel hierfür ist die Auswertung eines Selbsthilfeforums für Angehörige von inhaftierten Personen durch ein Forschungsteam der Technischen Hochschule Nürnberg und der Universität Passau. Im Rahmen des Projektes wurden 2881 Forumsbeiträge aus den Jahren 2005 bis 2020 mittels Topic Modelling – einem algorithmischen Verfahren zur automatischen Textanalyse – nach Diskussionsthemen sortiert (vgl. Ghanem et al. 2022, S.4). Anhand geclusterter Wörter, also Begriffe, die mit hoher Wahrscheinlichkeit bestimmte Oberthemen repräsentieren, können mit dieser Technik Themenschwerpunkte innerhalb des Forums identifiziert werden (vgl. ebd.).
Das Erstellen eines Ordnungssystems nach gewissen Oberbegriffen hilft den Forschenden ebenso dabei, „Themen für eine qualitative Analyse vorzustrukturieren.“ (ebd., S.5)

Durch den Gebrauch automatisierter Analyseverfahren lassen sich in der Sozialforschung komplett neue Möglichkeitsräume erschließen, die durch die Verwendung klassischer Methoden nicht oder nur unter sehr hohem Arbeitsaufwand erreicht werden könnten (vgl. ebd., S.4).

Bezüglich des Datenschutzes gibt die Forschungsmethode allerdings zu denken: Laut der Datenschutzgrundverordnung (DSGVO) besteht bei der Erhebung von Daten das Gebot der Datensparsamkeit und Zweckgebundenheit. Demnach müssen laut Art. 5 Abs. 1

„Personenbezogene Daten […]

b) für festgelegte, eindeutige und legitime Zwecke erhoben werden und dürfen nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden; […]

c) dem Zweck angemessen und erheblich sowie auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt sein („Datenminimierung“)“.

Diesem Gesetz steht das Prinzip von Big Data-Science konträr gegenüber, da es in der Natur von Big-Data liegt, dass sich genauere Zwecke der Datenanalyse erst im Nachhinein ergeben (vgl. ebd. S.5). Zunächst werden riesige Datenmengen pauschal heruntergeladen, da sich erst nach den Rechenprozessen ergibt, welche der gesammelten Daten eine Relevanz für das eigene Vorhaben besitzen (vgl. ebd.). Ebenso kann sich der Zweck der Untersuchung der Daten zum Beispiel durch Anwendung unüberwachter Lernverfahren verschieben, falls die KI unvorhergesehene Muster in Datensätzen findet, welche neue Forschungsfragen aufwerfen und dadurch neue Forschungszwecke prägen (vgl. ebd.). Ganz abgesehen davon sind die Daten in dem oben genannten Fall Gedanken betroffener Personen in einem zwar öffentlich zugänglichen Forum, welches allerdings zu Zwecken gegenseitiger Unterstützung von Personen mit Angehörigen in Haft gedacht war.

Mitarbeitenden des Forschungsprojekts ist an dieser Stelle geboten, Daten über Nutzende im Rahmen der wissenschaftlichen Arbeit so zu anonymisieren, dass Rückwärtssuchen beispielsweise über Zitieren von Beiträgen aus dem Forum verunmöglicht werden (vgl. ebd., S.7).

Ein nicht zu vermeidender Schaden könnte aber sein, dass die Aufmerksamkeit Dritter auf das Forum gelenkt und ein ursprünglicher Schutzraum so einer vermehrten öffentlichen Wahrnehmung ausgesetzt wird (vgl. ebd., S.8).

Aus diesem forschungsspezifischen Diskurs ergibt sich eine für Klient:innen relevante Fragestellung:

Wie sicher können sich Klient:innen Sozialer Arbeit sein, dass ursprünglich mit komplett anderen Absichten geteilte Informationen über die eigene Person nicht zweckentfremdet werden?
Ein Nachteil der zunehmenden Datafizierung menschlichen Lebens ist die Unübersichtlichkeit der Hinterlassung eigener Datenspuren im Netz. Dies kann marginalisierte Gruppen, die in intensiverem Austausch mit datenerhebenden Sozialbehörden stehen, besonders hart treffen.
So benennen Steiner und Tschopp das Problem des „function creeping“ als einen Effekt der zentralen Speicherung von Daten im Zuge der Digitalisierungsprozesse von Behörden (vgl. ebd. 2022, S.468). Dieses Phänomen beschreibt eine breitere Verwendung von Daten als zum Zeitpunkt der Erhebung der Daten deklariert, was in der Folge dazu führt, dass in gewissen Kontexten gesammelte Informationen genutzt werden, um in anderen Zusammenhängen die personenbezogenen Daten zur Kontrolle einzusetzen (vgl. Murakami Wood 2006, S.9).
So könnten bezogen auf das AFST Menschen mit psychischer Erkrankung, die das öffentliche Gesundheitssystem in Anspruch nehmen, Daten von sich preisgeben, über deren Verwendung bezüglich anderer Kontexte (Risikoscore) sie sich nicht im Klaren sind. Die ursprüngliche Funktion der personenbezogenen Daten (z.B. Anmeldung bei Hilfegruppe) würde dann auf eine Kontrollfunktion (Risikomodellierung AFST) erweitert.
Aufgrund digitaler Datenerhebung und algorithmischer Datenverarbeitung wird „function creeping“ also erstens zunehmend leichter und zweitens zunehmend attraktiver.
Doch für welche Verarbeitungsprozesse spielen personenbezogene Daten noch eine wichtige Rolle?

4.3 Predictive Policing (PP)

Werfen wir hierfür zunächst einen Blick auf das sogenannte Predictive Policing, zu Deutsch vorhersagende Polizeiarbeit.
Predictive Policing wird zur Prävention möglicher Straftaten eingesetzt, also „um eine noch nicht eingetretene Gefahr für ein Rechtsgut abzuwenden […].“ (Sommerer 2020, S.34) Hierbei wird auf algorithmengestützte Technologien zurückgegriffen (vgl. ebd.).
Predictive Policing funktioniert demnach analog zu den bereits vorgestellten Predictive Risk Modelling-Tools, beziehungsweise greift ebenso auf prädiktive Risikomodellierung zurück.

Der Begriff wird in die Kategorien ortsbezogenes Predictive Policing und personenbezogenes Predictive Policing geteilt.

Beim *ortsbezogenen Predictive Policing* bezieht sich die Vorhersage des Systems auf Orte, an denen kriminelles Verhalten zu bestimmten Zeiten wahrscheinlich sein wird (vgl. ebd., S.36). Das System soll der Polizei so eine effiziente strategische Planung in der Verteilung der Einsatzeinheiten ermöglichen (vgl. Rolfes 2017, S.52). Als Datengrundlage würden hier ortsbezogene Kriminalitätsstatistiken dienen, deren Auswertung auch mithilfe von kriminologischen Theorien erfolgen könnte (vgl. Sommerer 2020, S.36).

In dieser Arbeit stehen jedoch Risikomodelle, die sich nicht auf Orte, sondern Personen beziehen, im Vordergrund: Das *personenbezogene Predictive Policing* dient im weitesten Sinne dazu, Personen ein Risiko zuzuordnen, in Zukunft kriminell zu werden (vgl. ebd. S.37). Zur Berechnung des Risikos würde eine Analyse personenbezogener Daten von in der Vergangenheit kriminell gewordenen Personen als mustergebende Datengrundlage dienen (vgl. ebd.). Zur Berechnung der „Kriminalitätswahrscheinlichkeit" vergleicht der Algorithmus dann Eigenschaften der zu prüfenden Person mit denen der in der Vergangenheit straffällig gewordenen Menschen (vgl. ebd.). Je mehr Merkmale oder Verhaltensweisen der einzuschätzenden Person mit denen aus den Daten erkannten Mustern übereinstimmen, desto höher fällt die von der KI geschätzte Kriminalitätswahrscheinlichkeit aus.

Sommerer kritisiert an dieser Stelle, dass die Mustererkennung anhand von Parametern menschlicher Lebensführung zur Berechnung einer Kriminalitätswahrscheinlichkeit dem Erstellen von Persönlichkeitsprofilen gleiche, die dann mit kriminellem Verhalten gleichgesetzt würden (vgl. ebd., S.165). Die Drastik des Ganzen ergibt sich bei genauerem Nachdenken über diese Aussage:

Schränkt die Erstellung von Persönlichkeitsprofilen, die mit kriminellem Verhalten in Verbindung gebracht werden und darunter fallenden Leuten ein Verdachtsmoment zuordnen, nicht die freie Entfaltung der eigenen Persönlichkeit ein (vgl. ebd.)?

Schließlich gehen mit Verdächtigungen seitens der Polizei oftmals Kontrollelemente einher, was dadurch verstärkt wird, dass durch Anpassungen des Polizeigesetzes in deutschen Bundesländern die Schwelle für polizeiliche Überwachung enorm gesenkt wurde (vgl. Wyputta 2018). Die Prädiktoren, die in Algorithmen des Predictive Policing einfließen, sind jedoch keine Beweise für kriminelles Verhalten, sondern lediglich Angaben über verschiedene Arten menschlicher Lebensführung und -umstände, deren Verrechnung zu

einem Risikoscore immer nur eine Wahrscheinlichkeit darstellt und niemals zu einer sicheren Annahme führen kann. Nun würden gewisse Verhaltensweisen, die in ihrer puren Form erst einmal nichts mit Kriminalität zu tun haben, eine erhöhte Kontrollwahrscheinlichkeit nach sich ziehen; gewisse Lebensweisen wären also stigmatisiert.
Sommerer fügt an, dass größere Einschnitte in die Privatsphäre üblicherweise mit Erhöhung der Sicherheitsstandards – in diesem Fall bessere Prognosen durch die KI – gerechtfertigt werden könnten (vgl. ebd., S.165). Dies könnte hier zu einem sich selbst verstärkenden Durst nach Daten führen, da immer präzisere Vorhersagen nur durch immer mehr Datenquellen ermöglicht würden (vgl. ebd.).
Dieser Logik wohnt der Glaube inne, soziale Probleme wären durch technische Methoden lösbar, solange die dahinterstehenden Algorithmen nur präzise genug funktionierten und über eine dementsprechend große Datengrundlage verfügten. Die Annahme, psychosoziale/sozioökonomische Problemlagen wären technisch lösbar, nennt sich „Solutionism“ (vgl. Lehner 2020, S.141) und könnte kritischen Ansätzen, die sich mit struktureller Ursachenforschung beschäftigen, die Grundlage entziehen.

4.3.1 „HART“

Kommen wir nun einmal zu einem praktischen Anwendungsfall des PP, einem in Großbritannien entwickelten und eingesetzten Predictive Policing-Algorithmus, das sogenannte „Harm Assessment Risk Tool“ (HART) (vgl. Sommerer 2020, S78).
Bei diesem Tool handelt es sich um einen maschinell lernenden Algorithmus, welcher Polizeibeamt:innen dabei helfen soll, eine Entscheidung darüber zu fällen, ob in Gewahrsam genommene Personen bis zur Begutachtung dieser im Gerichtssaal in Gewahrsam behalten werden sollen (vgl. ebd). Die hier ausgegebene Prognose bezieht 34 aus kriminologischen Theorien abgeleitete Prädiktoren, zu denen u.A. Postleitzahl und Geschlecht zählen, in ihre Berechnungen mit ein und soll angeben, ob die in Gewahrsam genommene Person mit hoher, mittlerer oder niedriger Wahrscheinlichkeit innerhalb der nächsten zwei Jahre eine schwere Straftat begeht (vgl. Oswald et al. 2018, S.227-28). Als Datengrundlage wurde das Folgeverhalten von 104.000 in den Jahren 2008 bis 2012 festgenommenen Personen verwendet (vgl. ebd., S.228).
Zu HART sei noch zu erwähnen, dass die Programmierenden einem falsch-positivem Berechnungsfehler durch den Algorithmus weniger hohe Bedenken zumessen als einem falsch-negativem Fehler (vgl. ebd.). Das System wurde nämlich so justiert, dass es zwei Mal

häufiger vorkommt, dass eine eigentlich harmlose Person (keine Straftat in den zwei Folgejahren nach Festnahme) als gefährlich eingestuft wird, als dass eine gefährliche Person (schwere Straftat in den zwei Folgejahren) als ungefährlich eingestuft wird (vgl. ebd.).
Als schwere Straftaten werden zum Beispiel Mord und schwere Gewalttaten gelistet (vgl. ebd., S.227), was eine solche Gewichtung durchaus nachvollziehbar macht, allerdings lässt sich hieraus ableiten, dass das Gut der Risikovermeidung hier höher gewichtet wird als das Recht auf Bewegungsfreiheit.
Dieses Tool weist mehrere Ähnlichkeiten mit dem AFST auf, beispielsweise berechnet auch das HART die Gefahr des Eintritts eines als negativ bewerteten Ereignisses innerhalb der nächsten zwei Jahre. Die Anwendung nutzt außerdem ein breites Spektrum personenbezogener Daten zur Kalkulierung des Scores; auch deshalb merkt Sommerer kritisch an, dass die Anwendung einen Nachteil für von dem System als gefährlich klassifizierte Leute darstelle (vgl. ebd. 2020, S.79). Damit zusammenhängend kritisiert sie, dass durch das HART errechnete Vorhersagen öfter ein in Gewahrsam Behalten einer Person befürworteten als dies die Beamten täten (vgl. ebd.).
Auf die Aussage der Entwickler:innen, dass das Tool nur eine Entscheidungshilfe darstelle, entgegnet sie, dass es im Einzelfall betrachtet für eine Polizeikraft einer größeren Selbstsicherheit bedürfe, eine eigene Entscheidung gegen den Ratschlag des Algorithmus durchzusetzen als ohne äußeren Einfluss eine Entscheidung zu fällen (vgl. ebd.). Hiernach wären die Fachkräfte nach Sommerer in ihrer Entscheidungsfindung durch die KI eingeschränkt.
In der Fachliteratur gibt es einen Begriff, der das eben beschriebene Phänomen aufgreift: der sogenannte „automation bias“.
Dieser wird im Folgenden erklärt, bevor noch andere Kritikpunkte an algorithmischen Vorhersagen näher beleuchtet werden.

4.4 „Automation Bias“

„Automation bias“ ist ein englischer Fachbegriff, der vor allem zur Beschreibung menschlichen Verhaltens im Rahmen zunehmender Automatisierungsprozesse in Dienstleistungs- und Produktionsräumen an Bedeutung gewinnt (vgl. Mosier 1996, S.6-7). Gemeint ist mit dem Begriff die Zuschreibung erhöhter Kompetenz oder Fehlerfreiheit an eine Maschine (vgl. ebd., S.9-10). Dies kann zu fahrlässigem Verhalten seitens der die

Technik nutzenden menschlichen Fachkräfte führen, da diese oftmals höheres Vertrauen in eine scheinbar nicht fehlbare Technik als in ihre eigenen Lösungskompetenzen haben (vgl. ebd.). So wird schlussendlich ein von dem System ausgegebenes Ergebnis wenig hinterfragt, da sich stark darauf verlassen wird, dass die perfekte Lösung schon durch die Technologie gefunden wurde (Goddard 2012, S.121). In der Folge erodieren der Glaube an die eigene Einschätzung der Lage durch ein hochkomplex erscheinendes Entscheidungssystem, welches quantitative Grenzen der menschlich möglichen Informationsverarbeitung bei Weitem hinter sich lässt.

Besonders in stressigen Arbeitssettings bei der Bearbeitung komplizierter Aufgaben tritt dieser Effekt auf: So zeigten Studien, dass Arbeitende besonders bei einem hohen Arbeitsaufkommen zur schnelleren Bearbeitung eigener Aufgaben auf technische Hilfsmittel zurückgreifen (vgl. ebd., S.124).

Dies kann in Bezug auf die Nutzung algorithmischer Assistenzsysteme dazu führen, dass einer eigenen kritischen Reflexion im Vergleich zu einer scheinbar allwissenden KI keine besonders große Urteilsfähigkeit beigemessen wird (vgl. Gutwald et al. 2021, S.11).

Die Gefahr der Verdrängung menschlicher- durch algorithmischer Urteile wird im Folgenden Exkurs anhand von zwei Begriffen aus dem PP noch einmal genauer untersucht:

> *<u>Exkurs: Gefahrverdachtserzeugend vs. gefahrverdachtsbestätigend</u>*
>
> *Erinnern wir uns an dieser Stelle an die beschriebene Debatte zu Predictive Policing, in der Verfechter:innen des HART darauf beharrten, dass das Tool keinesfalls als eigenständiges Vorhersageinstrument fungiere, sondern lediglich zur Unterstützung der Polizeifachkräfte diene. Analoge Aussagen sind extrem typisch für die Entwickler:innen solcher Tools, da es eine ethisch sehr angreifbare Position wäre, moralische Entscheidungen technischen Systemen komplett zu überlassen. So finden sich ähnliche Aussagen auch von den Teams um KAIMo und dem AFST (vgl. Vaithianathan et al. 2017, S.4; Koska et al. 2023, S.12).*
>
> *Trotzdem ist es aufgrund emotionaler Belastungen, die den Entscheidungsträger:innen in sozial empfindlichen Bereichen zugemutet werden (vgl. Koska et al. 2023, S.11-12), denkbar, dass algorithmische Einschätzungen einen leichten Ausweg aus anspannungsreichen Abwägungsprozessen darstellen.*
>
> *Zurück zum Predictive Policing: Sommerer ordnet KI-Instrumente hier auf einer Skala ein, die eine Aussage über die Anwendungsbreite der KI preisgibt (vgl. ebd. 2020, S.38). So beginnt das Spektrum bei gefahrverdachtsbestätigenden Tools und endet bei*

gefahrverdachtserzeugenden Tools (vgl. ebd.). Gefahrverdachtsbestätigende Tools werden eingesetzt, um Risiken für Kriminalitätswahrscheinlichkeit bezüglich einzelner Personen abzuschätzen; gefahrverdachtserzeugende Tools erstellen Risikoscores für eine ganze Bandbreite an Personen wie zum Beispiel beim Fluggastdatenmusterabgleich, bei dem jährlich viele tausend Menschen ohne vorher bestehenden Verdacht pauschal von einer KI bewertet werden (vgl. ebd. S.96).

Die gefahrverdachtserzeugenden Tools stellen also einen stärker zu rechtfertigenden Eingriff in die Privatsphäre Einzelner dar, da hier Personen ohne konkrete Verdächtigung pauschal kontrolliert werden.

Technisch gesehen ist es leicht möglich aus einem gefahrverdachtsbestätigenden ein gefahrverdachtserzeugendes Tool zu machen, indem die Anzahl der durch das Tool geprüften Personen einfach erweitert wird (vgl. ebd., S.40).

So könnte sich auch im sozialen Bereich die Anwendungsbreite eines Predictive Risk Modelling Tools verwässern, wenn Fachkräfte aufgrund von Zeitdruck oder sozialem Stress das algorithmische Orakel befragen, obwohl eine Situation vorliegt, für die dies ursprünglich nicht geplant war. Denkbar wäre hier in Bezug auf KAIMo das zu Rate ziehen des Algorithmus ohne vorher abgesprochene Verdachtsmeldungen.

Eine Befürchtung, die das Wort „automation bias“ transportiert, ist das Vernachlässigen kritischer menschlicher Gedanken zugunsten von algorithmischer Entscheidungskompetenz. Allerdings entscheiden Algorithmen eben auf der Grundlage riesiger Datenmengen; befähigen die in den Trainingsdaten erfassten Erlebnisse die KI nicht zur Erlangung enormer Weisheit?
Und ist daher ein Verzicht auf Technik sogar ethisch fragwürdig, da ein Nicht-Nutzen der optimalen Lösung hohe Kosten mit sich bringen könnte?

4.5 Grenzen algorithmischer Möglichkeiten

Zur Beantwortung dieser beiden Fragen kann es lohnen, sich zunächst noch einmal die in den oberen Kapiteln beschriebene Funktionsweise der thematisierten KI-Systeme im Hinblick auf die Erstellung von Prognosen genauer vor Augen zu führen.
Nehmen wir das AFST als Beispiel, da hier die meisten Informationen über die genauen Datenquellen verfügbar sind: Das Tool wurde mit Daten von Fällen basierend auf knapp 77.000 Gefährdungsmeldungen trainiert (vgl. Vaithianathan 2017, S.11). Die hier eingespeisten Daten ordnen Lebensverhältnisse nach personenbezogenen Kategorien,

sodass vereinheitlichte Parameter entstehen, die von einem lernüberwachten Algorithmus leicht weiterverarbeitet werden können.
Zunächst ist darauf hinzuweisen, dass jegliche Information, die nicht in einen Parameter gepresst werden kann, in diesem Prozess verloren geht. Ebenso lässt der Algorithmus keinen Platz für Ambivalenzen, sodass beim AFST in der Berechnung des Risikoscores einer jeden Person in Bezug auf eine in der Vergangenheit liegende Anmeldung beim „behavioral health"-System die binären Klassifizierungsoptionen „Ja" und „Nein" bestehen. Jegliche Hintergrundinformation, zum Beispiel weshalb die Person das Hilfesystem in Anspruch nahm, findet keinen Platz innerhalb des Prädiktors.
So könnte es, um die Thematik etwas zu veranschaulichen, bei einer Person durch formellen Zwang aufgrund einer gerichtlichen Entscheidung dazu kommen, dass eine Einweisung in eine suchttherapeutische Einrichtung erfolgt, während sich eine andere Person aus freien Stücken dazu entscheidet, Hilfe in Anspruch zu nehmen. Diese zwei komplett unterschiedlichen Motive würden durch den Rechenprozess des Algorithmus aus der Bewertung der Person herausfallen und in der Konsequenz einheitlich betrachtet werden (vgl. Lehner 2020, S.140).
Zurück zur Kerntätigkeit der Berechnungen eines KI-Systems: Dieses versucht, aus Daten der Vergangenheit Informationen über die Zukunft zu extrahieren; denn so hoch technisiert wie die Verfahren auch sein mögen, ihre Datengrundlage kann sich immer nur auf vergangene Ereignisse stützen. Eine KI stellt durch ihre Kalkulation Zusammenhänge zwischen Prädiktoren und Ausgabewerten her, die am Ende zum Beispiel durch die in Risikoscores enthaltenen Wahrscheinlichkeitsberechnungen ihre Form erhalten. An dieser Stelle ist es wichtig zu erkennen, dass sich der Risikoscore zwar auf einen Menschen bezieht, diesen jedoch nur anhand der Verhaltenszusammenhänge der Lebenswelten anderer Personen beurteilen kann. Auch hier hilft ein Beispiel, diesen Mechanismus umfänglicher zu verstehen: Bleiben wir bei dem Prädiktor „behavioral health" des AFST. Der Risikoscore des Falles ist 0-3 Punkte höher, wenn der:die mutmaßliche Täter:in in der Vergangenheit auf das verhaltensgesundheitliche Hilfesystem zurückgegriffen hat (vgl. Gerchick et al. 2023). Der Algorithmus hat hier also einen Zusammenhang zwischen der Registrierung im öffentlichen „behavioral health"-System und einer möglichen Kindeswohlgefährdung erkannt. Dieser Zusammenhang ist nicht „unwahr", allerdings sollte einem bewusst sein, dass die Vorhersage der KI durch Verrechnung diverser psychosozialer und sozioökonomischer

Einflussfaktoren aus den vergangenen Fällen zu Stande kam.
Stellen wir uns einmal eine Person vor, die kurz nach ihrer Schwangerschaft eine postpartale Depression erleidet, sich über diesen Umstand bewusst wird und unter Anderem zum Schutz ihres Kindes professionelle Hilfe sucht. Diese Person würde vom Algorithmus mit hoher Wahrscheinlichkeit einen höheren Risikoscore zugeordnet bekommen als eine Mutter, die unter selben Umständen keine Hilfe holt. Hieran wird deutlich, dass die errechneten Zusammenhänge nur in Bezug auf gesellschaftliche Durchschnittswerte als wirklich valide gelten können.
Das AFST modelliert durch die Verrechnung verschiedener unabhängiger Variablen also ein Bild von gesellschaftlicher Vergangenheit, welches im Anschluss dazu dienen soll, eine Prognose für zukünftige Einzelschicksale zu erstellen.
Lehner kritisiert an der Verwendung von PRM in Sozialer Arbeit, dass „[eine] kritisch reflektierte Sozialarbeit [darauf bedacht sein sollte], zu verhindern, dass die algorithmisch gesponnenen Schicksalsfäden sich in Ketten verwandeln.“ (ebd. S.140)
Hiermit greift er die durch Nutzung algorithmischer Prognoseverfahren entstehende Gefahr der Perpetuierung vergangener Verhältnisse auf, welche durch den Effekt des „automation bias“ Gefahr laufen, verstärkt zu werden.
Auch dieses Zitat wird zur Veranschaulichung noch einmal auf das obige Beispiel bezogen: In diesem Fall würde die gesellschaftliche Zuschreibung, dass psychische Krankheit mit Kindeswohlgefährdung oder Vernachlässigung zusammenhängt, in den Köpfen Sozialarbeitender durch höhere Risikoscores der betroffenen Personen verstetigt. Hier ist allerdings die Frage, ob die Stärkung dieses gesellschaftlichen Stereotyps nicht erst dazu beiträgt, Einsichtsfindung (und damit auch Inanspruchnahme von Hilfen) bei betroffenen Elternteilen aus Angst vor Verurteilung durch Andere zu behindern.
An dieser Stelle könnte jedoch ebenso eingewandt werden, dass der AFST-Risikoscore nur Callscreenern angezeigt wird, keine Begründung für das Zustandekommen des Scores enthält und im Optimalfall zur Entdeckung und Inanspruchnahme einer passenden Hilfe führt.
Es gibt bei der Anwendung des AFST also durchaus ein Für und Wider, allerdings kann auf die oben gestellten Frage entgegnet werden, dass die Formulierung „enorme Weisheit“ einerseits übertrieben ist, da die Prognosealgorithmen vorklassifizierte Informationen nutzen, weshalb nur eine quantifizierte Realitätsdarstellung möglich ist, die

Qualitätseinbußen zur Folge hat. Und dass andererseits Qualitätseinbußen auch zur Entstehung von „Kosten“ beitragen können.

4.6 Bias

Qualitätseinbußen können auch durch die in den Daten versteckten Einstellungen derjenigen, die die Trainingsdaten erheben oder Vorgehensweisen vorgeben, nach denen Daten gesammelt werden, verursacht werden (vgl. Lehmann et al. 2021, S.24). In Bezug auf den Aufbau des Algorithmus und die Erhebung von Trainingsdaten stellen sich auch Fragen wie, wer bestimmt, welche Daten Relevanz besitzen? Durch die Wahrnehmungsprozesse welcher Personen werden Daten gefiltert? Oder, in welchen Kontexten wurden Daten gesammelt?
Hier wäre ein praktisches Beispiel, dass zu verschriftlichende Kataloge zum Status des Kindeswohls in deutschen Jugendämtern aufgrund von Zeitmangel oftmals erst nach dem Fällen einer Entscheidung ausgefüllt werden (vgl. Koska et al. 2023, S.8). Deshalb würden essentielle Informationen zum Fall hier so angepasst, dass die gefällte Entscheidung im Nachhinein legitim erscheint (vgl. ebd.).
Durch solche Prozesse fließen schiefe Bilder der Realität in Trainingsdaten ein, die Bias genannt werden.

> „Ein Bias bezeichnet allgemein Verzerrungseffekte. Die Psychologie versteht darunter Einstellungen oder Stereotypen, welche die Wahrnehmung unserer Umwelt, Entscheidungen und Handlungen positiv oder negativ beeinflussen. Diese Beeinflussung kann unbewusst (impliziter Bias) oder bewusst (expliziter Bias) geschehen. In der Statistik wird ein Bias als Fehler im Rahmen der Datenerhebung und -verarbeitung […] verstanden" (Beck 2019, S.8)

Neben den oben bereits beschriebenen Möglichkeiten, wie Bias Einzug in Daten erhält, können Verzerrungen ebenso durch unvollständige Datensätze oder mangelnde Verfügbarkeit gewisser Daten entstehen (vgl. Heesen et al. 2021, S.130).
Der Ursprung dieser Verzerrungen liegt oftmals in den Daten, die in den Trainingsprozess eingeflossen sind, was in Bezug auf PRM schwere Auswirkungen haben kann:
Das Thema „Bias“ hängt nämlich stark mit Diskriminierung zusammen, da stereotype Menschenbilder sich beispielsweise in Dokumentationen von Fachkräften des Jugendamtes befinden können, die wiederum als Datengrundlage für algorithmische Vorhersagen dienen sollen.

Innerhalb der Trainingsdaten der von uns untersuchten Algorithmen könnten sich also verzerrte Wahrnehmungen von Personengruppen verbergen, die zu einer unfairen Gewichtung algorithmischer Prädiktoren führen (vgl. vhb o.J.).
Betrachten wir diesen Umstand vor dem Hintergrund, dass seitens der Fachkräfte, die Tendenz besteht, algorithmischen Assistenzsystemen aufgrund der Verarbeitung riesiger Datenmengen eine fast schon transzendente Urteilsfähigkeit zuzuschreiben („automation bias"), wäre die Verstetigung bestehender Diskriminierungsmuster möglich, da Sozialarbeitenden durch die Einführung solcher Systeme der Anlass zur selbständigen Reflexion entzogen werden könnte.

4.7 KI und Diskriminierung

Gibt es einen Schutz vor Diskriminierung durch KI und was ist der Unterschied von Diskriminierung zu den Unterscheidungen aufgrund algorithmischer Kategorien?
Hierzu kann es helfen, das Wort „Diskriminierung" zunächst einmal zu definieren:

> „Der Begriff Diskriminierung bedeutet in seiner lateinischen Herkunft, aber auch teils im Englischen und in fachsprachlichen Kontexten ethisch indifferent ›unterscheiden, absondern, trennen‹." (Heesen et al. 2021, S.129)

In diesem Kontext wird der Begriff also moralisch neutral als ein Unterscheidungsprozess definiert ohne den ein Algorithmus, der stets mit klassifiziererten Operatoren arbeitet, schlicht nicht auskommen würde (vgl. ebd, S.131).
Der Unterschied zu der Bedeutung des Begriffs, die die meisten Menschen aus alltäglichen Kontexten mit dem Wort assoziieren, liegt in der Frage nach Gerechtigkeit:
Gibt es eine gerechtfertigte Grundlage für die Unterscheidung bzw. stehen die Klassifizierungen in Zusammenhang mit „[herabsetzenden] oder [benachteiligenden] [Absichten] gegen Angehörige bestimmter sozialer Gruppen […]." (Hormel et al. 2010, S.7)
Nach dieser Definition erzeugt Diskriminierung also eine ungerechtfertigte Ungleichbehandlung für Angehörige sozial kategorisierbarer Gruppen.
Beck et al. zeichnen die Schwelle für negativ konnotierte Diskriminierung im Rahmen algorithmischer Datenverarbeitung in zwei unterschiedlichen Dimensionen:
Von Diskriminierung nach verbreitetem Verständnis kann erstens gesprochen werden, wenn Menschengruppen nach bereits existierenden stereotypen Bildern klassifiziert werden oder

wenn diese zweitens nach anderen Kriterien eingeteilt, aber dennoch ungerechtfertigt systematisch benachteiligt werden (vgl. ebd. 2019, S.12).
Diskriminierungen sind in Deutschland nach §§1,2 AGG (Allgemeines Gleichbehandlungsgesetz) weitestgehend untersagt. Auch in den USA gibt es mehrere sogenannte „federal laws“ – Gesetze, die im kompletten Land gelten -, die diverse Diskriminierungsformen in verschiedenen Kontexten verbieten (vgl. Civil Rights Division 2023).
In der Fachliteratur werden Diskriminierungen nicht nur anhand betroffener Personengruppen kategorisiert, sondern auch nach der ihr innenwohnenden Funktionsweise.
Vier dieser Diskriminierungsformen besitzen in Bezug auf algorithmische Datenverarbeitung besondere Relevanz und werden daher im Folgenden kurz thematisiert:

- *Direkte Diskriminierung*: Hierunter wird die am leichtesten sichtbare Form der Diskriminierung verstanden. Es ist ein direkter Zusammenhang zwischen einer Gruppe und der die Gruppe benachteiligenden Handlungs- oder Verhaltensweisen erkennbar, wobei auch ein Nicht-Handeln diskriminierend sein kann (vgl. Heesen et al. 2021, S.132).
- *Indirekte Diskriminierung* funktioniert weniger explizit, da sie sich nicht auf die diskriminierte Gruppe in spezifischer Form bezieht, sondern die Gruppe mittelbar betrifft (vgl. Hormel et al. 2010, S.30). So könnte bezogen auf das AFST (V1) der Einbau sozioökonomischer Faktoren schwarze Bevölkerung indirekt diskriminieren, da diese in der Historie der USA und bis zum heutigen Tage aufgrund von Unterdrückungsstrukturen tendenziell schlechteren Zugang zu finanziellem Kapital haben (vgl. Gerchick et al. 2023).
- *Statistische Diskriminierung* erfolgt ebenso wie die indirekte Diskriminierung auf mittelbare Art und Weise, wird jedoch als spezifische Kategorie für diskriminierende Annahmen verwendet, die aufgrund einer statistischen Korrelation, die in diesem Fall einen Zusammenhang suggeriert, der nicht sicher vorhanden ist, entstehen (vgl. Heesen et al. 2021, S.132). Als Beispiel hierfür nennen Heesen et al. die Verweigerung eines Kredits für Personen aus Wohngegenden, in denen Kredite in der Vergangenheit des häufigeren nicht gezahlt wurden (vgl. ebd.).
- *Emergente Diskriminierung* ist eine sehr technische Form der Diskriminierung und entsteht durch Wechselwirkungen zwischen KI, eingeflossenen Trainingsdaten, sich verändernder Realität und Interpretationen der Anwender:innen (vgl. Beck et al. 2019, S.9). Emergenter

Bias stellt eine neue und schwer identifizierbare Form der Diskriminierung dar, da sie durch ein hochkomplexes Zusammenspiel technischer und menschlicher Komponenten entsteht (vgl. Heesen 2021, S.130).
Diese letzte Form der Diskriminierung wirft auch Fragen bezüglich unserer Leitfrage auf:
Wie kann die Soziale Arbeit im gesellschaftlichen Diskurs Stellung bezüglich möglicher eigens produzierter Diskriminierungen durch KI-gestützte Assistenzsysteme beziehen, wenn diese für die Arbeitenden selbst immer schwerer zu durchschauen sind?
Wie kann generell Verantwortung für Entscheidungen übernommen werden, wenn einem das Zustande Kommen dieser nicht ganz klar ist?

4.8 Blackbox-Problematik

Wie bereits unter dem Punkt „2.5.1 Künstliche neuronale Netzwerke (KNNs)“ demonstriert bestehen maschinelle Lernverfahren zu Teilen aus einer zwar rational funktionierenden, sich dennoch menschlicher Nachvollziehbarkeit entziehenden Blackbox. Der Grund für die große Intransparenz solcher technischen Methoden ist dem maschinellen Lernen, bei dem Programmierende keine detaillierten Problemlösestrategien entwerfen, sondern die Software selbst mit einer „Lernarchitektur“ ausstatten, inhärent (vgl. Martini 2019, S.42). Besonders Algorithmen die auf Verfahren des Deep-Learnings beruhen, weisen eine große Intransparenz auf (vgl. Lehmann 2021, S.21):
So bauen KNNs durch Fütterung mit Trainingsdaten ihre eigene Struktur zwar nur im Rahmen vorgegebener Möglichkeiten aus, verfügen allerdings am Ende eines Trainingsprozesses über viele tausend automatisch ausdefinierte Interaktionsbeziehungen zwischen den Neuronen.
Hieraus folgt, dass bei Anwendungen algorithmischer Assistenzsysteme in sensiblen sozialen Bereichen wichtige diskursive Prozesse wegfallen könnten, wenn sich zunehmend auf KI-gestützte Urteile verlassen wird.
Um diese Schlussfolgerung noch einmal in Perspektive zu setzen, malen Heese et al. das Bild noch ein wenig weiter aus und stellen bei zunehmender Abbildung öffentlicher Meinungen durch algorithmische Datenerhebungen statt durch intersubjektives Handeln einen Verfall diskursiver Öffentlichkeit in Aussicht (vgl. ebd. 2021, S. 143-144). Gemeint ist damit in etwa, dass, wenn algorithmische Systeme zunehmend die Darstellung gesellschaftlicher Realitäten übernehmen, sich dies wiederum auf die Bildung von Gesellschaft auswirken wird.

Genährt wird die Gefahr eines diskursiven Verfalls im menschlichen Zusammenleben durch die Blackbox-Problematik:
Aufgrund der eigenständigen Ausbildung komplexer Systeme - siehe die autonome Justierung der Aktivitätsbeziehungen zwischen den Neuronen eines KNNs - ist es extrem schwierig, Ergebnisse maschinell lernender Anwendungen für menschliche Logiken nachvollziehbar darzustellen (vgl. Martini 2019, S.43). Im Fachdiskurs wird hier von fehlenden Rekonstruktionsmöglichkeiten in Bezug auf die von KI erzeugten Ergebnisse gesprochen (vgl. ebd).
Die versteckten Neuronenschichten stehen hier sinnbildlich für die Problematik, da das künstliche neuronale Netzwerk in diesem Bereich des Systems Abstraktionsprozesse erlernt, die wir nur in Teilen verstehen können.
Aufgrund der Algorithmisierung gesellschaftssteuernder Schnittstellen wie zum Beispiel durch die Einführung KI-gestützter Assistenzsysteme in der Institution Jugendamt entsteht die Gefahr der Diskursverarmung:
Es könnten sich Funktionsteile des gesellschaftlichen Systems hier also Diskussionen entziehen, wenn Diskriminierungsmuster durch den Blackbox-Charakter mitentscheidender künstlicher Intelligenzen erstens weniger nachvollziehbar werden und sich zweitens die Übernahme von Verantwortung schwieriger gestaltet als zuvor.
Denn nach unseren bisherigen Erkenntnissen besteht ein Widerspruch zwischen dem wissenschaftlich nachgewiesenen „automation bias“ und der Aussage von KI-System-Herstellenden, dass endgültige Entscheidungen nur von Menschen getroffen werden könnten, obwohl alle bis hierhin vorgestellten Systeme zumindest eine implizite Aussage über berechnete Entscheidungspräferenzen ausgeben. Die Diskrepanz zwischen denjenigen, die Systeme entwickeln oder für deren Einführung an Arbeitsplätzen sorgen, und denjenigen, die mit algorithmischen Assistenzsystemen arbeiten, könnte in dem Sinn dafür liegen, wer letzten Endes die Verantwortung übernimmt, wenn eine algorithmisch ausgegebene Entscheidung negative Konsequenzen nach sich zieht – Entwickler:in oder Anwender:in? Reißt die Einführung von KI-Systemen also eine Lücke in das kollektive Verantwortungsbewusstsein von Firmen oder Organisationen?

Exkurs: Entscheidungsbeeinflussung oder Gehirnerweiterung?

Auch das Projektteam hinter KAIMo, welches den Algorithmus eher als eine

Reflexionshilfe als eine entscheidungsverändernde Komponente der Arbeit im Jugendamt vorstellt, hat mit dem Projekt das Ziel, Prozesse während der Kindeswohlgefährdungseinschätzung effizienter zu gestalten. Daraus folgt womöglich, dass bei gleichem Output (Entscheidung bezüglich Kindeswohlgefährdung) weniger menschliche Arbeitszeit verwendet werden soll. Der Reflexionsprozess, der zur Verantwortungsbildung beiträgt, wird jedoch nicht weniger arbeitsintensiv und ist folglich in Teilen der Maschine überlassen.
Es könnte demnach auch argumentiert werden, dass aufgrund zukünftig weniger eingeplanter menschlicher Arbeitszeit ab einem gewissen Punkt im Arbeitsprozess auf eine KI-gestützte Einschätzung vertraut werden müsste. So wäre weniger Einblick in Entscheidungsfindungsprozesse allein schon aufgrund von Zeitmangel denkbar, der Algorithmus würde auch nach diesem Modell mit menschlicher Entscheidungshoheit konkurrieren.

Demgegenüber steht eine philosophische Theorie von David Chalmers und Andy Clark, die sich mit der Frage beschäftigen, wo sich der Übergang zwischen eigener Kognition und dem Rest der Welt befindet (vgl. ebd. 1998, S.7). Hierbei schreiben die Autoren der Umwelt eines Menschen eine aktive Rolle bei gedanklichen Prozessen zu (vgl. ebd.). Die Beispiele für das Zusammenspiel von Gehirn und externer Welt liegen hier auf einer Skala von der Zuhilfenahme eigener Finger bei Rechenaufgaben über Taschenrechner bis hin zu Bildschirmen, die uns helfen, Formen von Objekten zu visualisieren (vgl. ebd. S. 7-11).
Die beiden Autoren geben in diesem Kontext ein Beispiel von 2 Personen - Otto und Inga -, die ins Museum gehen wollen. Otto hat Alzheimer, informiert sich deshalb zunächst im eigenen Notizbuch über die Adresse, während Inga auf ihre Erinnerung an den Straßennamen vertraut (vgl. ebd. S.12-13). Aus diesem Beispiel kann geschlossen werden, dass kognitive Funktionen, die unser Bild der Realität maßgeblich beeinflussen, nicht zwangsweise als Gedanken in ihrer puren Form in unserem Kopf existieren, sondern oft mit Gegenständen aus unserer Umwelt in Verknüpfung funktionieren (vgl. ebd. S.13).

So könnte auch in Bezug auf Assistenzsysteme wie KAIMo argumentiert werden, dass der Algorithmus als gehirnerweiternde Orientierung in Denkprozessen der Jugendamtsmitarbeitenden fungiert.
Ein entscheidender Unterschied zu Taschenrechner oder Notizblock besteht allerdings darin, dass KAIMo sich als „moral agent" an flexiblen Denkprozessen beteiligen soll und daher in aktiverer Wechselwirkung mit dem:der Anwender:in steht als ein analoges

Speichermedium oder eine Rechenhilfe. Die Einflussnahme findet außerdem auf einem uneindeutigeren Spektrum statt und ist schwer zu rekonstruieren, was uns wieder zu einer unklarer geregelten Übernahme von Verantwortung führt.

Heinrichs et al. beschreiben einen Lösungsweg zum Schließen dieser Verantwortungslücke: So gehen sogenannte „techno-optimists" davon aus, dass sich diese Lücke durch Kontrolle der KI-Systeme schließen lasse (vgl. ebd. 2022, S.44). Eine Möglichkeit zur Verantwortungszuschreibung bestehe darin, dass Entwickler:innen der KI-Systeme Erklärungszusammenhänge für die Berechnungen der hauseigenen Algorithmen darlegen könnten (vgl. ebd.).

4.9 Whitebox und Blackboxtest

Zur Überprüfung der Zusammensetzung KI-gestützter Urteile kann sich allerdings nicht immer auf Herausgeber:innen der Systeme verlassen werden, sodass auch unabhängig von der den Algorithmus entwerfenden Partei Testmöglichkeiten entwickelt wurden.
Eine der Kontrollmöglichkeiten nennt sich „Whitebox-Test" (vgl. Martini 2019, S.45). Der Name des Tests spielt auf die Transparenz der Entwicklerfirma bezüglich grundlegender Annahmen und Mechanismen des Algorithmus an, die es erleichtern, Zusammenhänge zwischen abhängigen und unabhängigen Variablen offenzulegen (vgl. ebd.).
Deutlich undurchlässiger wird das Testergebnis, wenn Informationen über die Grundarchitektur der KI wegfallen, sodass ohne jegliche Vorkenntnisse, die Ergebniskonstruktion des Algorithmus geprüft werden muss.
Eine Prüfung ohne Kenntnis informatischer Strukturen wird „Blackbox-Test" genannt und greift auf Werkzeuge des „reverse engineering", zu Deutsch etwa umgekehrte Konstruktion, zurück (vgl. ebd.). Mithilfe dieser Werkzeuge soll vor allem erkannt werden, ob das KI-System diskriminierende Auswirkungen zur Folge hat.

> „Sie ziehen aus der statistischen Auswertung einer Vielzahl einzelner Verarbeitungsvorgänge Rückschlüsse darauf, wie der Quellcode aufgebaut ist und nach welcher Struktur deterministische Algorithmen zu ihren Ergebnissen kommen." (Martini 2019, S.45)

Als Datengrundlage für Blackbox-Tests dienen also Verarbeitungsvorgänge, von denen zwar keine Details wie die Gewichtung einzelner Variablen bekannt sind, jedoch haben diejenigen, die die Tests durchführen, Einsicht in eingegebene Parameter sowie in die dazugehörigen Ausgabedaten (vgl. ebd.).
Sobald sich jedoch keine Informationen über die verwendeten Einflussfaktoren erörtern lassen, stößt auch diese Methode an ihre Grenzen.
Würde das Team hinter KAIMo zum Beispiel keine Informationen darüber veröffentlichen, welche Daten, in die Analysen des Assessment-Bots einfließen und ebenso ausgegebene Reflexionsfragen, Beurteilungen oder Akzentuierungen der jeweiligen Merkmalslage unter Verschluss halten, wäre es so gut wie unmöglich, Einsicht in algorithmische Ergebnisfindung zu erhalten.
Allerdings ist dies kein realistisches Szenario, da das Jugendamt von öffentlichen Geldern finanziert wird, weshalb Transparenz an dieser Stelle entscheidend für Akzeptanz und Förderung des Projekts ist.
Es besteht dementsprechend ein Eigeninteresse der Entwickler:innen, eine KI zu entwerfen, in der eine Ergebnisherleitung von Anfang an integriert ist.
Koska verweist diesbezüglich auf den Forschungsbereich der „Explainable AI“, welcher sich zum Ziel gemacht hat, algorithmisch produzierte Ausgabewerte besser erklärbar zu machen.

4.10 XAI – „Explainable AI“

Algorithmische Vorhersagen werden durch rationale Prozesse generiert. Was durch den Quellcode in Bezug auf Predictive Risk Modelling in mathematischen Vektoren abgebildet werden soll, berührt jedoch Logiken, deren Herleitung auf eine völlig andere Art und Weise zustande kam als durch die in algorithmischen Trainingsprozessen ausgearbeitete Funktionsweise von KI-Technologien.
Aufgrund der unterschiedlichen Herangehensweise an die logische Konstruktion sozialer Zusammenhänge, kann es zu Schwierigkeiten kommen, wenn an der Schnittstelle von Mensch und Maschine Kausalitäten anders abgebildet werden (vgl. Holzinger et al. 2020, S.37).
Auch für den geplanten Einsatz von Chatbots im Rahmen des KAIMo-Projekts ist es von zentraler Bedeutung, dass Ergebnisse des Risikomodells während eines Austauschs zwischen Mensch und Bot erklärbar und hinterfragbar bleiben.

Dies macht die Auseinandersetzung mit „Explainable AI" - im Fachdiskurs zumeist „Interpretable AI“ genannt (vgl. Adadi et al. 2018, S. 52142) – zu einem wichtigen Forschungsfeld, sollte KAIMo in Zukunft Anwendung in deutschen Jugendämtern finden. Auch in anderen Fachbereichen gilt, dass algorithmische Vorhersagen ohne eine dazugehörige Erklärung über das Zustandekommen Selbiger einen deutlich niedrigeren Wert für Anwender:innen haben, da Interventionsmöglichkeiten erst richtig sortiert werden können, nachdem Problemzusammenhänge kausal strukturiert wurden (vgl. Cohausz 2022, S.361).

Doch wie weit ist dieser Forschungsbereich fortgeschritten? Gibt es bereits Methoden, die KI-Entscheidungen nachvollziehbarer machen können?

Erstens ist festzustellen, dass erklärbare KI ein junges, jedoch schnell wachsendes Forschungsgebiet darstellt, innerhalb dessen noch viele unbetretene Terrains existieren (vgl. Cohausz 2022, S.361; Adadi et al. 2018; S.52138).

So entstehen immer noch verschiedenste Ansätze, die sich zum Ziel setzen, Problemlösungswege und Entscheidungsverhalten künstlicher Intelligenz für menschliche Logiken nachvollziehbar zu machen.

Um zu veranschaulichen, wie maschinell generierte Ergebnisse interpretierbar gemacht werden können, folgt nun beispielhaft die Erklärung einer Methode aus der Praxis:

Mark T. Keane und Barry Smyth widmen sich in einer wissenschaftlichen Arbeit aus dem Jahr 2020 einem XAI-Ansatz, der versucht, sich KI-Entscheidungsfindungswegen über die Analyse des algorithmischen Umgangs von scharf kontrastierbaren Fällen zu nähern (vgl. ebd. 2020, S.1).

4.10.1 „Counterfactual explanations“

Bevor diese kompliziert klingende Methode genauer beleuchtet wird, nähern wir uns dieser Erklärungsstrategie genau wie die beiden Autoren selbst zunächst aus psychologischer Perspektive und nehmen das Thema Kindeswohlgefährdung als Beispiel:

So fordert die US-Amerikanische Rechtsanwältin Robin Frank in Bezug auf das AFST die Offenlegung des Risikoscores für betroffene Familien sowie eine bessere Nachvollziehbarkeit des Zustandekommens des Selbigen (vgl. Arias 2023; Ho et al. 2022). Als Erklärungshilfe, weshalb die eigene Familie einen hohen Risikoscore zugeteilt bekam, könnte hier ein anonymisierter, ähnlicher Fall aus dem Datensatz des Algorithmus herangezogen werden, bei dem eine Kindeswohlgefährdung vorlag.

Dies würde die Entscheidung des Algorithmus zwar schwerer angreifbar machen, allerdings noch keine befriedigenden Erkenntnisse darüber liefern, was in der Familie anders hätte laufen müssen, um nicht in das Raster des Algorithmus zu fallen (vgl. Keane et al. 2020, S.2). An dieser Stelle könnten die sogenannten „counterfactual explanations“, zu Deutsch kontrafaktische Erklärungen, zum Einsatz kommen:
Wäre es in diesem Fall nicht wissenswerter, weshalb der Algorithmus bei einem immer noch vergleichbarem Fall zu einer anderen Einschätzung, also zu einem niedrigeren Risikoscore kam? Welche Einflussfaktoren (unabhängige Variablen) müssten sich verändern, damit der Risikoscore eines Falles sinkt?
Keane und Smyth argumentieren, dass kontrafaktische Erklärungen am ehesten der menschlichen Intuition entsprechen würden und untermauern ihren Punkt mit Forschungsergebnissen, die darauf hindeuten, dass ein solcher Erklärungsansatz aus psychologischer Perspektive sinnvoller erscheint (vgl. ebd. 2020, S.2-3). Jedenfalls liefert dieser konkrete Ansatzpunkte für das Nachvollziehen einer algorithmischen Entscheidung.

Um herauszufinden, wie sich gewisse Einflussfaktoren oder Kombinationen aus Einflussfaktoren auf konkrete Berechnungen auswirken, können Fallpaare aus sogenannten „nearest-unlike-neighbors (NUN)“, zu Deutsch nächste ungleiche Nachbarn, gebildet werden (vgl. ebd., S.2).
Diese Fallpaare weisen möglichst wenige unterschiedliche Eigenschaften auf, werden jedoch von der KI ungleich klassifiziert (vgl. ebd., S.5). Wenn ein Algorithmus zwischen zwei ähnlichen Fällen durch eine ungleiche Klassifizierung der Fälle differenziert, ermöglicht der Vergleich beider Fälle eine Eingrenzung bei der Suche nach für die Entscheidung maßgeblichen Prädiktoren. Durch diese Vorgehensweise erschließen sich das Ergebnis beeinflussende Unterschiede in der Merkmalslage eines Falles und der:die Anwender:in erhält bessere Einsicht in die Mechanismen der Berechnung.
Es werden also durch die Betrachtung feiner Nuancen und Kontraste von Fallpaaren entscheidende Einflussfaktoren ermittelt.
Bei näherem Betrachten fällt auf, dass der Ansatz der Entwickler:innen von KAIMo, die angeben, ihr eigenes Endprodukt hänge noch von der technologischen Entwicklung im Bereich von XAI ab, sich sehr an dem Prinzip der „counterfactual explanations“ orientiert: So hat der Planning-Bot die Aufgabe, entscheidende Merkmale eines Falles herauszuarbeiten, anstatt eine Prognose zu erstellen. Ähnlich funktioniert die Herangehensweise der

„counterfactual explanations“, die eine möglichst große Übereinstimmung der Einflussfaktoren bei dennoch unterschiedlichen Ergebnissen berechnen sollen. So werden ebenfalls entscheidende Eingabeparameter herausgearbeitet - im übertragenen Sinne also Fallmerkmale.
„Counterfactual explanations“ werden – wie sich fast schon vermuten lässt – in der Praxis nicht von Hand sondern durch bestimmte Algorithmen erstellt, die sogenannten Interpretationsalgorithmen. Dieser Term beschreibt eine Oberkategorie an Algorithmen, an welche die essenzielle Anforderung gestellt wird, Vorhersagen der ursprünglichen algorithmischen Modelle in für Menschen verständlichen Konzepten zu erklären (vgl. Li et al. 2022, S.3201).
Zur Prüfung der Qualität dieser speziellen Algorithmen kann ein sogenannter „sanity check“ durchgeführt werden, welcher feststellt, ob die Ergebnisse des Interpretationsalgorithmus‘ auch vertrauenswürdig sind (vgl. ebd., S.3200).
Es existieren mittlerweile diverse Interpretationsalgorithmen (vgl. ebd., S.3210), sodass davon ausgegangen werden kann, dass diese auch als Reaktion auf sich erhöhenden gesellschaftlichen und politischen Druck (vgl. Kaur et al. 2022, S.2-4) zukünftig mehr Anwendung finden werden.
Unter Berücksichtigung der Tatsache, dass Chatbots mittlerweile in der Lage dazu sind, auch komplexe Konzepte semantisch sinnvoll zu strukturieren, kann hier also die Hoffnung bestehen, dass eine KI ihre eigenen Entscheidungen in einer reflektierten Art und Weise begründen können wird.

5. Fazit

Innerhalb dieser wissenschaftlichen Arbeit wurde sich mit Hinblick auf die in der sozialen Arbeit notwendige Reflexionsarbeit mit unterschiedlichsten Konzepten um das Thema „KI im sozialen Bereich“ befasst.
Ein zentraler Unterschied zwischen den zwei näher beleuchteten Projekten aus den USA und Deutschland besteht in dem Ziel des Teams um KAIMo ein algorithmisches Vorhersagetool zu entwickeln, welches anstatt einer quantifizierten Ausgabe in Form eines Risikoscores auch in der Lage sein soll, pädagogische Verfahren qualitativ zu unterstützen. Dieser Ansatz suggeriert zunächst eine Förderung reflexiver Prozesse.
Eine Idee der Entwickler:innen ist, KAIMo im Sinne des „fleet learnings“ in mehreren Jugendämtern zu installieren, um die KI zunächst passiv im Schattenmodus dazulernen zu

lassen.
Hierbei ergeben sich auf den ersten Blick datenschutzrechtlichen Bedenken, da personenbezogene Daten nicht direkt für den Zweck der konkreten Fallbearbeitung durch das zuständige Jugendamt verwendet würden, sondern in einer zentralen Datenbank zu Trainingszwecken eines Algorithmus gespeichert wären. Diesbezüglich müsste man sich hier mit der Gefahr des „function creeping“ auseinandersetzen, die aber durch eine strenge Anonymisierung von Falldaten vermeidbar wäre.
Ebenso könnten die hohen Ansprüche des Teams an ihre KI zu einem hohen Durst an Daten führen: Im sozialen Bereich ist der Erfolg einer Hilfe zu großen Teilen von Beziehungsarbeit und anderen sozialdynamischen Vorgängen abhängig, deren Abbildung allein schon aufgrund von Zeitmangel in ihrem vollen Umfang nicht annähernd in die Dokumentation der Fachkräfte einfließt.
Folgendes würde zurzeit zwar niemals von Entwickler:innen einer solchen KI geäußert werden, es ist jedoch nicht auszuschließen, dass für zukünftige Algorithmen Datenquellen wie z.B. die Audioaufnahmen von sozialen Beratungsgesprächen etc. als zusätzlich zu bisher bestehenden Datengrundlagen herhalten müssten, um präzisere Kalkulationen zu ermöglichen.
Da dies jedoch aktuell keinen Platz in den Zukunftsvisionen des Teams um KAIMo findet, bleiben nur Dokumentationen der Fachkräfte als Datenquelle übrig. Ob diese ausreichen, um menschliche Reflexionsprozesse zu fördern bleibt abzuwarten, in Bezug auf den Vergleich zum „fleet learning“ muss jedoch festgestellt werden, dass Entscheidungen Autofahrender im Straßenverkehr deutlich besser in Daten zu fassen scheinen als soziale Reflexionsprozesse. Auch aufgrund der Tatsache, dass der Straßenverkehr klar definierte Regeln kennt, die in Bezug auf den Umgang mit Personen schwieriger zu fassen sind, hinkt diese Analogie an manchen Stellen.
Allerdings formuliert das KAIMo-Team als Vorgabe für die eigene KI, sie solle eher als Reflexionsstütze gelten und nicht als autonom agierende Prozessoreinheit. Je nach dem wie direkt Handlungsvorschläge der KI auf einer Skala von offenen Fragen bis hin zum konkreten Vorschlagen von Maßnahmen ausdefinieren wird, kann hier durchaus Potenzial für Reflexionsunterstützungen gesehen werden.
Im Kontrast zu KAIMo reduziert sich der Anwendungsbereich des AFST auf ein konkretes Arbeitsfeld des DHS. Eine möglichst schnelle Einschätzung der Gefahrensituation soll hier

durch die Ausgabe eines Risikoscores unterstützt werden. Inwiefern die Arbeit in diesem Bereich vor der Einführung des AFST Gegenstand von Reflexionsprozessen war, ist schwierig abzuschätzen, jedoch bleibt aufgrund von zeitlichen Vorgaben und Mangel an Kontakt mit betroffenen Klient:innen der Behörde weniger Raum für eine ausgewogene Betrachtung sozialer Umstände.

Dies negiert jedoch nicht, dass auch hier Entscheidungen über menschliches Leben getroffen werden. In diesem Arbeitsfeld besitzen Urteile der Mitarbeitenden ebenso eine große Tragweite in Bezug auf Verteilung menschlicher Lebenschancen, womit die Arbeit an GPS-Fallmeldungen mit der Ausübung struktureller Macht und somit Verantwortungsübernahme verbunden sind.

Es ist davon auszugehen, dass durch die Einführung des AFST aufgrund des „automation bias‘“ ein Teil menschlicher Entscheidungskompetenz an den Algorithmus übergegangen ist, welcher eigene Urteile nicht selbst reflektiert.

Das DHS argumentiert hier, dass in dem AFST enthaltener Bias zwar vorhanden sei, sich allerdings gesellschaftliche Stigma im Vergleich zu dem Zeitraum vor der Einführung des Tools reduzierten (vgl. Allegheny County DHS 2019, S. 10).

Kritiker:innen des Tools, die verschiedene Möglichkeiten zur Analyse des Tools haben (Whitebox-Test, Blackbox-Test, Einholung von Daten) beklagen auf der anderen Seite im Vergleich zum gesellschaftlichen Durchschnitt unproportionale Entscheidungen zu Ungunsten gewisser sozialer Gruppen (vgl. Gerchick 2023). So entspinnt sich in Bezug auf ein algorithmisches PRM-Tool eine gesellschaftliche Debatte, in der integrierte Prädiktoren diskutiert werden.

Es ist zu erwähnen, dass die Machtverteilung einer solchen Debatte nicht immer symmetrisch ist, da Wissen über die Funktionsweise von PRM-Algorithmen ungleich verteilt sein kann. Interessant für unsere Fragestellung ist jedoch die Beobachtung, dass die Einführung von KI-Instrumenten im sozialen Bereich Diskussions- und Reflexionsprozesse in eine breitere gesellschaftliche Schicht schwappen lässt.

Auf der anderen Seite sind die Versuche, das eigene Tool gesellschaftlich transparent zu machen, jedoch nicht in die Entscheidung gemündet, einen Fokus auf die Einführung eines Interpretationsalgorithmus zu legen. So bleiben Entscheidungen auch für Mitarbeitende wohl weiter in Teilen undurchschaubar, was in der Konsequenz eine Auseinandersetzung mit kritischen Aspekten wie dem Anteil von Bias verunmöglicht.

Des Weiteren eröffnet das Zusammendenken der Begriffe gefahrverdachtserzeugend, gefahrverdachtserweiternd und „automation bias“ die Sicht auf das Risiko einer überbordenden Anwendung algorithmischer Technologien im sozialen Bereich. So könnten Reflexionsprozesse Mitarbeitender in sozialen Bereichen verarmen, wenn KI-Systeme über den ursprünglich angedachten Anwendungsbereich hinaus benutzt würden.
Ein weiterer Themenbereich ist die Spannung zwischen Reflexionsprozessen und dem Ordnen komplexer Lebensverhältnisse mittels quantifizierter Parameter. Das AFST arbeitet hier mit sehr binären Kategorien, wodurch sich die Gefahr der Entstehung blinder Flecken bei der Beurteilung von Subjekten ergibt.
In Bezug auf KAIMo kann gespannt in die Zukunft geblickt werden, um besser darüber urteilen zu können, ob die Wahl von Textdateien als Arbeitsgrundlage des Algorithmus dazu in der Lage ist, Lebensrealitäten pluraler darzustellen.
Eine geregelte Verantwortungsübernahme von Entscheidungen, die durch Algorithmen mit Blackbox-Charakter beeinflusst wurde, hängt maßgeblich an der Etablierung einer klaren Verteilung von Verantwortung für bestimmte Szenarien (Wer ist verantwortlich? Entwickler:innen vs. Anwender:innen) sowie dem Einpflegen von Interpretationsalgorithmen und deren technische Fähigkeiten.
Wenn also eine KI-Software Interpretationsalgorithmen sowie eine zuverlässig funktionierende Chatbot-Funktion vereinen würde, könnte dies die Akzeptanz und Nachvollziehbarkeit in Bezug auf algorithmische Prozesse auch für Laien ermöglichen.
Ebenso ist ein Aufklären der Fachkräfte über die Aussagekraft algorithmischer Vorhersagequalität sowie deren Grenzen von zentraler Bedeutung für die weitere Entwicklung in diesem Bereich.

II. Literaturverzeichnis

-3Blue1Brown (2017): But what is a neural network? | Chapter 1, Deep learning [YouTube]. Online unter: https://www.youtube.com/watch?v=aircAruvnKk (Zugriff: 11.06.2023).

-Adadi, Amina/ Berrada, Mohammed (2018): Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). In: *IEEE Access*, (Vol. 6), S. 52138-52160.

-Alle, Frederike (2012): Kindeswohlgefährdung. Das Praxishandbuch. Aufl. 2. Freiburg im Breisgau: Lambertus-Verlag.

- Allegheny County (a) (2023): The Allegheny Family Screening Tool. Predictive Risk Modeling in Child Welfare in Allegheny County. [Website] Online unter: https://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx (Zugriff: 21.07.2023).

Allegheny County (b) (2023): Behavioral health. [Website] Online unter: https://www.alleghenycounty.us/Human-Services/About/Offices/Behavioral-Health.aspx (Zugriff: 19.09.2023)

- Allegheny County DHS (2019): Section 7. Frequently-Asked-Questions. Online unter: https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/FAQs-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-8.pdf (Zugriff: 18.09.2023).

- Allegheny County DHS (2023): Welcome to Allegheny Connect. A place to find ressources and support in Allegheny County. [Website] Online unter: https://connect.alleghenycounty.us/ (Zugriff: 18.09.2023).

- Alpaydin, Ethem (2022): Maschinelles Lernen. 3. Aufl. Berlin/Boston: Walter de Gruyter GmbH.

- Annie E. Casey Foundation (2021): Frequently Asked Questions About Juvenile Probation. Online Unter: https://www.aecf.org/blog/frequently-asked-questions-about-juvenile-probation (Zugriff: 23.06.23).

- Arias, Vicky (2023): Parents challenge how A.I. software is used in child welfare cases. In FISM News. Online unter: https://fism.tv/parents-challenge-how-a-i-software-is-used-in-child-welfare-cases/ (Zugriff: 13.09.23).

- Beck, Susanne/ Grunwald, Armin/ Jacob, Kai/ Matzner, Tobias (2019): Künstliche Intelligenz und Diskriminierung. Herausforderungen und Lösungsansätze. In: Plattform Lernende Systeme (Hrsg.). Online unter https://www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-

loesungsansaetze.html (Zugriff: 18.09.2023).
- Bode, Beatrice (2023): Chat GPT: US-Magazin lässt Artikel seit Monaten heimlich von KI schreiben. Online unter: https://www.basicthinking.de/blog/2023/01/13/cnet-laesst-ki-seit-wochen-finanz-artikel-schreiben/ (Zugriff 20.07.2023).
- Burghardt, Jennifer (o.J.): Kapitel 3, Lektion 3. Einführung in KAIMo Projekt [Audio-Interview]. In: virtuelle hochschule bayern (Hrsg.): KI und Soziale Arbeit [Online-Kurs]. Online unter: https://open.vhb.org/course/view.php?id=236&chapter=3&selected_week=18 (Zugriff: 19.07.2023).
- Chowdhary, K. R. (2020). *Fundamentals of Artificial Intelligence.*
- *Civil Rights Division (2023): Federal Protections Against National Origin Discrimination. Online unter: https://www.justice.gov/crt/federal-protections-against-national-origin-discrimination-1#ed (Zugriff: 24.07.2023).*
- *Clark, Andy/ Chalmers, David (1998): The extended mind. In: Analysis (Vol. 58, Nr.1), S. 7-19.*
- *Cohausz, Lea (2022):* Towards Real Interpretability of Student Success Prediction Combining Methods of XAI and Social Science. In: Proceedings of the 15th International Conference on Educational Data Mining, 361–367. Online unter: https://zenodo.org/record/6853069 (Zugriff: 08.09.2023).
- Dalton, Erin (2022): Response to AP article „An algorithm that screens for child neglect raises concerns". Online unter: file:///C:/Users/nicol/Downloads/DHS%20response%20to%20AP%20article_Algorithm%20that%20screens%20for%20child%20neglect-3.pdf (Zugriff: 18.09.2023).
- *Gapski, Harald (2020), Digitale Transformation. Datafizierung und Algorithmisierung von Lebens- und Arbeitswelten. In: Kutscher, Nadia/ Ley, Thomas/ Seelmeyer, Udo/ Siller, Friederike/ Tillmann, Angela/ Zorn, Isabel (Hrsg.): Handbuch Soziale Arbeit und Digitalisierung. Weinheim: Beltz, 156–166.*
- *Gerchick, Marissa/ Jegede, Tobi/ Shah, Tarak/ Guitiérrez, Ana/ Beiers, Sophie/ Shemtov, Noam/ Xu, Kath/ Samant, Anjana/ Horowitz, Aaron (2023):* The Devil is in the Details: Interrogating Values Embedded in the Allegheny Family Screening Tool. Online unter: https://www.aclu.org/the-devil-is-in-the-details-interrogating-values-embedded-in-the-allegheny-family-screening-tool (Zugriff: 15.07.2023).
- *Ghanem, Christian/ Eckl, Markus/ Lehmann Robert (2022): Big Data und Forschungsethik in der Sozialen Arbeit. In EthikJournal (8. Jg.│Ausgabe 1/2022).*

- *Goddard, Kate/ Roudsari, Abdul/ Wyatt, Jeremy (2012): Automation bias: a systematic review of frequency, effect mediators, and mitigators. In:* Journal of the American Medical Informatics Association (2012;19), S.121-127.
- *Google (2023): Datenschutzerklärung. Online unter https://policies.google.com/privacy?hl=de&fg=1&utm_source=ucbs#infocollect (Zugriff: 10.07.2023).*
- *Gutwald, Rebecca/ Burghardt, Jennifer/ Kraus, Maximilian/ Reder, Michael/ Lehmann, Robert/ Müller, Nicolas (2021): Soziale Konflikte und Digitalisierung. Chancenund und Risiken digitaler Technologien bei der Einschätzung von Kindeswohlgefährdungen. In: EthikJournal (7. Jg.│Ausgabe 2/2021).*
- *Gutwald, Rebecca* (o.J.): Kapitel 3, Lektion 3. Einführung in KAIMo Projekt [Audio-Interview]. In: virtuelle hochschule bayern (Hrsg.): KI und Soziale Arbeit [Online-Kurs]. Online unter: https://open.vhb.org/course/view.php?id=236&chapter=3&selected_week=18 (Zugriff: 19.07.2023).
- Hasan, Ragib /Dutta, Amit Kumar (2013): How Much Does Storage Really Cost? Towards a Full Cost Accounting Model for Data Storage. In: Altmann, Jörn/ Vanmechelen, Kurt/ Rana, Omer F. (Hrsg.): Economics of Grids, Clouds, Systems, and Services. GECON 2013. Lecture Notes in Computer Science (Vol. 8193), S.29-43. Cham: Springer International Publishing.
-Heesen, Jessica/ Reinhardt, Karoline/ Schelenz, Laura (2021): Diskriminierung durch Algorithmen vermeiden: Analysen und Instrumente für eine demokratische digitale Gesellschaft. In: Bauer, Gero/ Kechaja, Maria/ Engelmann, Sebastian/ Haug, Lean (Hrsg.): *Diskriminierung und Antidiskriminierung: Beiträge aus Wissenschaft und Praxis*. Bielefeld: transcript Verlag, S.129-148.
- Heinrichs, Bert/ Heinrichs, Jan-Hendrik/ Rüther, Markus (2022): Künstliche Intelligenz. In: Birnbacher, Dieter/ Stekeler-Weithofer, Pirmin/ Tetens, Holm (Hrsg.): Grundthemen Philosophie. Berlin/Boston: Walter de Gruyter GmbH.
- Hinne, Max/ Gronau, Quentin F./ van den Bergh, Don/ Wagenmakers, Eric-Jan (2020): Advances in Methods and Practices in Psychological Science, 200-215. Online unter: https://journals.sagepub.com/doi/full/10.1177/2515245919898657 (Zugriff: 07.07.23).
- Heurer, Klaus (o.J.): Sokratische Methode (5./4. Jahrhundert v. Chr.). Online unter: https://die-bonn.de/zeitzeichen/sokratischemethode (Zugriff: 07.07.2023).
- Ho, Sally/ Burke, Garrance (2022): An algorithm that screens for child neglect in Allegheny

County raises concerns. In: WESA. Online unter: https://www.wesa.fm/politics-government/2022-04-29/an-algorithm-that-screens-for-child-neglect-in-allegheny-county-raises-concerns (Zugriff: 13.09.2023).
- Holzinger, Andreas/ Heimo, Müller (2020): Verbinden von Natürlicher und Künstlicher Intelligenz: eine experimentelle Testumgebung für Explainable AI (xAI). In: HMD Praxis der Wirtschaftsinformatik (Vol. 57), S.33-45.
- Hormel, Ulrike/ Scherr, Albert (2010): Diskriminierung als gesellschaftliches Phänomen. In: Hormel, Ulrike/ Scherr, Albert (Hrsg.): Diskriminierung. Grundlagen und Forschungsergebnisse. 1. Aufl. Wiesbaden: VS Verlag.
-*IBM (a) (o.J.):* What is a knowledge graph?. Online unter: https://www.ibm.com/topics/knowledge-graph (Zugriff: 03.07.2023).
- IBM (b) (o.J.): Was ist starke KI?. Online unter: https://www.ibm.com/de-de/topics/strong-ai (Zugriff 20.07.2023).
- iRights.Lab (2017): Das Recht auf informationale Selbstbestimmung. Online unter: https://www.bpb.de/themen/recht-justiz/persoenlichkeitsrechte/244837/das-recht-auf-informationelle-selbstbestimmung/ (Zugriff: 10.07.2023).
- Jaskolla, Ludwig (o.J.): Kapitel 3, Lektion 3. Einführung in KAIMo Projekt [Audio-Interview]. In: virtuelle hochschule bayern (Hrsg.): KI und Soziale Arbeit [Online-Kurs]. Online unter: https://open.vhb.org/course/view.php?id=236&chapter=3&selected_week=18 (Zugriff: 19.07.2023).
- Kasten, Hartmut (2014): Entwicklungspsychologische Grundlagen der frühen Kindheit und frühpädagogische Konsequenzen. In KiTaFachtexte. Online unter: https://www.kita-fachtexte.de/fileadmin/Redaktion/Publikationen/KiTaFT_kasten_2014.pdf (Zugriff: 18.09.2023).
- Kaur, Davinder/ Uslu, Suleyman/ Rittichier, Kaley J./Durresi, Arjan (2022): Trustworthy Artificial Intelligence: A Review. In: ACM Computing Surveys (Vol. 55, Issue 2, Artikelnr.: 39), S.1-38. Online unter: https://dl.acm.org/doi/10.1145/3491209 (Zugriff: 16.09.23).
- Keane, Mark T./ Smyth, Barry (2020): Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). In: arXiv (2005.13997). Online unter: https://arxiv.org/abs/2005.13997 (Zugriff 13.09.2023).
- Kirste, Moritz/ Schürholz, Markus (2019): Einleitung: Entwicklungswege zur KI. In: Wittpahl, Volker (Hrsg.): Künstliche Intelligenz. Technologie|Anwendung|Gesellschaft. Wiesbaden:

Springer Verlag. Online unter: https://www.researchgate.net/publication/330051036_Einleitung_Entwicklungswege_zur_KI_Technologie_Anwendung_Gesellschaft/link/63e703b56425237563a4441b/download (Zugriff: 20.07.2023).
- Koehrsen, Will (2018): Introduction to Baysesian Linear Regression. An explanation of the Bayesian approach to linear modeling. Online unter: https://towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7 (Zugriff: 07.07.2023).
- Koska, Christopher/ Reder, Michael (2023): KI-gestützte Assistenz für moralische Konfliktsituationen. Zur Algorithmisierung im Handlungsfeld der Kindeswohlgefährdung. In: Trost, Kai Erik (Hrsg): Der Wert der Freundschaft in der mediatisierten Alltagswelt. Eine narratologisch-semiotische Analyse der Freundschaftserzählungen Jugendlicher. Medienethik Digitale Ethik (Band 19). Stuttgart: Franz Steiner Verlag. [Zum Zeitpunkt der Abgabe noch nicht veröffentlicht, Veröffentlichung voraussichtlich im Oktober 2023]
- Kraus, Maximilian (o.J.): Kapitel 3, Lektion 3. Einführung in KAIMo Projekt [Audio-Interview]. In: virtuelle hochschule bayern (Hrsg.): KI und Soziale Arbeit [Online-Kurs]. Online unter: https://open.vhb.org/course/view.php?id=236&chapter=3&selected_week=18 (Zugriff: 19.07.2023).
- Lehmann, Robert/ Albrecht, Jens/ Domes, Michael/ Petrlic, Ronald/ Bradl, Marion/ Burghardt, Jennifer/ Kiener, Dagmar/ Stieler, Mara/ Widerhold, Jean-Pierre/ Zauter, Sigrid (2021): Gutachten über die Einsatzmöglichkeiten von Künstlicher-Intelligenz-Software in aufsuchenden, digitalen Angeboten der Migrationsberatung. Fem.OS/Technische Hochschule Nürnberg (Hrsg.). Online unter: https://minor-kontor.de/kuenstliche-intelligenz-in-der-migrationsberatung/ (Zugriff: 19.07.2023).
-Lehner, Nikolaus (2020): Digitale Technologie zwischen Überwachung, sozialer Kontrolle und Fürsorge. In: Kutscher, Nadia/ Ley, Thomas/ Seelmeyer, Udo/ Siller, Friederike/ Tillmann, Angela/ Zorn, Isabel (Hrsg.) Handbuch Soziale Arbeit und Digitalisierung 1.Aufl. Weinheim: Beltz Juventa Verlag., S.129-144.
- Lenzen, Manuela (2020): Künstliche Intelligenz: Fakten, Chancen, Risiken. München: C.H. Beck.
- Li, Xuhong/ Xiong, Haoyi/ Li, Xingjian/ Wu, Xuanyu/ Zhang, Xiao/ Liu, Ji/ Bian, Jiang/ Dou, Dejing (2022): Interpretable deep learning: interpretation, interpretability, trustworthiness,

and beyond. In: *Knowledge and Information Systems (Vol.* **64)**, 3197–3234. Online unter: https://link.springer.com/article/10.1007/s10115-022-01756-8#citeas (Zugriff: 16.09.2023).
-Lohaus, Arnold (2018): Entwicklungspsychologie des Jugendalters. Berlin: Springer Verlag.
- Martini, Mario (2019): Blackbox Algorithmus – Grundfragen einer Regulierung künstlicher Intelligenz. Berlin/Heidelberg: Springer Verlag. Online unter: https://link.springer.com/book/10.1007/978-3-662-59010-2 (Zugriff 21.07.2023).
- Mayer, Christoph P. (2018): Künstliche Intelligenz und Maschinelles Lernen: Hintergrund, Anwendungsfälle und Chancen für Medienunternehmen. In: MedienWirtschaft (Jahrgang 15, Heft 3). Hamburg: New Business Verlag.
- Mosier, Kathleen/ Skitka, Linda (1996): Human Decision Makers and Automated Decisions Aids: Made for Each Other? In: Parasuraman, Raja/ Mouloua, Mustapha (Hrsg.): Automation and Human Performance: Theory and Applications (pp. 201–220). New Jersey: Lawrence Erlbaum Associates. Online unter: https://www.researchgate.net/publication/230601064_Human_Decision_Makers_and_Automated_Decision_Aids_Made_for_Each_Other (Zugriff 14.07.2023).
- Murakami Wood, David (2006): A Report on the Surveillance Society. For the Information Commissioner by the Surveillance Studies Network. London, Information Commissioner's Office.
- Müller, Burkhard (2018): Eingriff. In: Otto, Hans-Uwe/ Thiersch, Hans/ Treptow, Rainer/ Ziegler, Holger (Hrsg.): Handbuch Soziale Arbeit. Grundlagen der Sozialarbeit und Sozialpädagogik. 6 Auflage. München: Ernst Reinhardt Verlag.
- Nam, Jinseok/ Kim, Jungi/ Loza Mencía, Eneldo/ Gurevych, Iryna/ Fürnkranz, Johannes (2014): Large-Scale Multi-label Text Classification — Revisiting Neural Networks. In: Calders, Toon/ Esposito, Floriana/Hüllermeier, Eyke/ Meo, Rosa (Hrsg.): Lecture Notes in Computer Science, (Vol. 8725), 437–452. Berlin: Springer Verlag. Online unter https://link.springer.com/chapter/10.1007/978-3-662-44851-9_28 (Zugriff: 03.07.2023).
- Ng, Annalyn/Soo, Kenneth (2018): Data Science – was ist das eigentlich?. Algorithmen des maschinellen Lernens verständlich erklärt. Berlin: Springer Verlag. Online unter: https://link.springer.com/book/10.1007/978-3-662-56776-0 (Zugriff 10.07.2023)
- OECD (2023): Poverty rate. Online unter: https://data.oecd.org/inequality/poverty-rate.htm (Zugriff: 21.06.2023).
- Oswald, Marion/ Grace, Jamie/ Urwin, Sheena/ Barnes, Geoffrey (2018): Algorithmic risk

assessment policing models: Lessons from the Durham HART model and “Experimental” proportionality. In: Information & Communications Technology Law (Vol. 27, 2018 – Issue 2), 223-250.
- Paaß, Gerhard/ Hecker, Dirk (2020): Künstliche Intelligenz. Was steckt hinter der Technologie der Zukunft?. Wiesbaden: Springer Verlag.
- Reder, Michael: Kann ein Algorithmus moralisch kalkulieren?. KAIMo. [Website] Online unter: https://www.kaimo.bayern/ (Zugriff: 21.09.2023).
- Rizvi, Munaza Batool/ Conners, Gregory P./ King, Kevin C./ Lopez, Richard A./ Bohlen, Julie/ Rabiner, Joni (2023): Pennsylvania Child Abuse Recognition and Reporting. In: Statpearls [Internet]. Online unter: https://www.ncbi.nlm.nih.gov/books/NBK565852/ (Zugriff: 18.09.2023).
- Rolfes, Manfred (2017): Predictive Policing: Beobachtungen und Reflexionen zur Einführung und Etablierung einer vorhersagenden Polizeiarbeit. In: Universität Potsdam (Hrsg.): Potsdamer Geographische Praxis (Nr.12), S.51-76.
- Rubin, Jonathan (2020): Children, Youth & Families Bulletin. Number 3490-20.08. Online unter: https://www.dhs.pa.gov/docs/Publications/Documents/FORMS%20AND%20PUBS%20OCYF/OCYF%20Bulletin%203490-19-02%20Statewide%20General%20Protective%20Services%20GP%20Referrals_12202019.pdf (Zugriff: 19.08.2023).
- Sommerer, Lucia (2020): Personenbezogenes Predictive Policing. Kriminalwissenschaftliche Untersuchung über die Automatisierung der Kriminalprognose. 1. Aufl. Baden-Baden: Nomos.
- Steiner, Oliver/ Tschopp, Dominik (2022): Künstliche Intelligenz in der Sozialen Arbeit. Grundlagen, Entwicklungen, Herausforderungen. In: Sozial Extra 46, 466-471. Online unter: https://link.springer.com/article/10.1007/s12054-022-00546-4 (Zugriff: 20.07.2023).
- Stoetzer, Matthias W. (2017): Regressionsanalyse in der empirischen Wirtschafts- und Sozialforschung Band 1. Eine nichtmathematische Einführung mit SPSS und Stata. Berlin: Springer Verlag. Online unter: https://link.springer.com/book/10.1007/978-3-662-53824-1 (Zugriff: 21.07.2023).
- Strobl, Chris (2017): Update: Tesla’s Fleet Learning. How the Silicon Valles software company will win the autonomous car war. Online unter:

https://blog.hackerbay.com/update-teslas-fleet-learning-8e34c3cd6ab4 (Zugriff: 07.07.2023).
- Vaithianathan, Rhema/ Putnam-Hornstein, Emily/ Jiang, Nan/ Nand, Parma/ Maloney, Tim (2017): Developing Predictive Models to Support Child Maltreatment Hotline Screening Decisions: Allegheny County Methodology and Implementation. Online unter: https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V1-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL.pdf (Zugriff: 18.09.2023).
- Vaithianathan, Rhema/ Kulick, Emily/ Putnam-Hornstein, Emily/ Benavides-Prado, Diana (2019): Allegheny Family Screening Tool: Methodology, Version 2. Online unter: https://www.alleghenycountyanalytics.us/wp-content/uploads/2019/05/Methodology-V2-from-16-ACDHS-26_PredictiveRisk_Package_050119_FINAL-7.pdf (Zugriff: 18.09.2023).
- Vaithianathan, Rhema/ Putnam-Hornstein, Emily/ Chouldechova, Alexandra/ Benavides-Prado, Diana/ Berger, Rachel (2020): Hospital Injury Encounters of Children Identified by a Predictive Risk Model for Screening Child Maltreatment Referrals. Evidence From the Allegheny Family Screening Tool. In: Jama Pediatrics (Vol.174, Nr.11). Online unter: https://www.researchgate.net/publication/343413164_Hospital_Injury_Encounters_of_Children_Identified_by_a_Predictive_Risk_Model_for_Screening_Child_Maltreatment_Referrals_Evidence_From_the_Allegheny_Family_Screening_Tool/link/600f1fbc299bf14088c06ad5/download (Zugriff: 18.09.2023).
- Vela, Daniel/ Sharp, Andrew/ Zhang, Richard/ Nguyen, Trang/ Hoang, An/ Pianykh, Oleg S. (2022): Temporal quality degradation in AI models. In: Scientific Reports (Vol. 12). Online unter: https://www.nature.com/articles/s41598-022-15245-z (Zugriff: 20.07.2023).
- vhb (virtuelle hochschule bayern) (o.J.): Kapitel 1, Lektion 4. Ergebnisse und Output von KI [Audio-Interview]. In: virtuelle hochschule bayern (Hrsg.): KI und Soziale Arbeit [Online-Kurs]. Online unter: https://open.vhb.org/course/view.php?id=236&chapter=1&selected_week=8 (Zugriff: 19.07.2023).
- Wyputta, Andreas (2018): Verschärftes Polizeigesetz in NRW. Verdächtig sind alle, die so aussehen. In: taz. Online unter: https://taz.de/Verschaerftes-Polizeigesetz-in-NRW/!5499063/ (Zugriff 22.07.2023).
-Zhao, Tianna/ Zhang, Yuanjian/ Zhang, Hongyun/ Miao, Duoqian (2023): Multi-granular labels with three-way decisions for multi-label classification. In: International Journal of Machine Learning and Cybernetics. Ohne Seitenzahlen. Online unter:

https://link.springer.com/article/10.1007/s13042-023-01861-2#citeas (Zugriff: 03.07.2023).

- Zuboff, Shoshana (2019): Surveillanca Capitalism – Überwachungskapitalismus. In: APuZ (69. Jahrgang, 24-26), 4–9.

III. Eidesstattliche Erklärung

Ich versichere, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind in allen Fällen unter Angabe der Quelle kenntlich gemacht.

Hamburg, 25.09.2023 Unterschrift